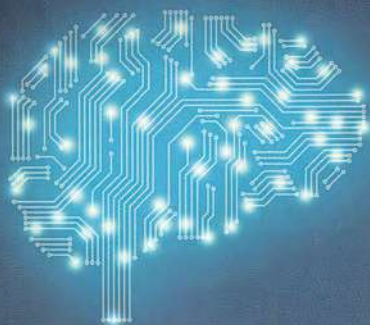


„Un ghid pasionant pentru provocările și alegerile aduse de căutarea unui viitor măreț al vieții, inteligenței și conștiinței – pe Pământ și dincolo de el.“ —ELON MUSK

MAX TEGMARK

VIATA 3.0



OMUL
ÎN EPOCA
INTELIGENȚEI
ARTIFICIALE

HUMANITAS



Una dintre cele mai spectaculoase evoluții în cursul celor 13,8 miliarde de ani scurși de la big bang este că acea materie mută și neînsuflețită a devenit inteligentă. Cum s-a petrecut asta și cum poate progresa inteligența în viitor? Ce are de spus știința despre istoria și destinul inteligenței în cosmosul nostru? —**MAX TEGMARK**

VIAȚA 3.0

Max Tegmark (n. 1967) este un fizician american de origine suedeză care și-a manifestat încă din adolescență talentul de informatician, scriind în cod mașină un procesor de text pe care l-a vândut corporației suedeze Dataindustrial AB. După ce a obținut licența în fizică la Institutul Tehnologic Regal din Stockholm, a studiat în continuare fizica la Universitatea Berkeley din California, unde și-a susținut doctoratul. În prezent, este profesor la Institutul Tehnologic din Massachusetts (MIT). Cercetările sale din domeniul cosmologiei l-au condus către „ipoteza universului matematic” (prezentată în cartea sa *Our Mathematical Universe*), conform căreia toate structurile care există din punct de vedere matematic există și fizic. Preocupat de inteligența artificială și îngrijorat de perspectivele dezvoltării ei, Max Tegmark a înființat Institutul pentru Viitorul Vieții, prin care a reușit să aducă în centrul atenției specialiștilor și publicului larg probleme delicate legate de siguranța inteligenței artificiale.

MAX TEGMARK

VIATA 3.0

OMUL ÎN EPOCA INTELIGENȚEI ARTIFICIALE

Traducere din engleză
de Vlad Zografi



HUMANITAS
BUCUREȘTI

Coperta: Ioana Nedelcu
Tehnoredactor: Manuela Măxineanu
Corectori: Cristina Jelescu, Alexandru Anghel
DTP: Florina Vasiliu, Dan Dulgheru

Tipărit la Paper Print – Brăila

Max Tegmark
Life 3.0: Being Human in the Age of Artificial Intelligence
Copyright © 2017 by Max Tegmark
All rights reserved.

© HUMANITAS, 2019, pentru prezenta versiune românească

Descrierea CIP a Bibliotecii Naționale a României
Tegmark, Max
Viața 3.0: Omul în epoca inteligenței artificiale / Max Tegmark;
trad. din engleză de Vlad Zografi. – București: Humanitas, 2019
ISBN 978-973-50-6402-0
I. Zografi, Vlad (trad.)
004.8

EDITURA HUMANITAS
Piața Presei Libere 1, 013701 București, România
tel. 021/408 83 50, fax 021/408 83 51
www.humanitas.ro

Comenzi online: www.libhumanitas.ro
Comenzi prin e-mail: vanzari@libhumanitas.ro
Comenzi telefonice: 021/311 23 30

*Pentru Echipa FLI,
care a făcut totul posibil*

Mulțumiri

Le sunt sincer îndatorat tuturor celor care m-au încurajat și m-au ajutat să scriu această carte, între care:

familia mea, prietenii, profesorii, colegii și colaboratorii, pentru sprijinul și inspirația pe care mi le-au oferit de-a lungul anilor, Mama, pentru că mi-a stârnit curiozitatea în privința conștiinței și a sensului,

Tata, pentru spiritul de luptător pus în slujba unei lumi mai bune, fiii mei Philip și Alexander, pentru că m-au făcut să văd cum se naște minunea inteligenței umane,

toți pasionații de știință și tehnologie de pretutindeni, care, de-a lungul anilor, prin întrebări și comentarii m-au încurajat să-mi urmez și să-mi public ideile,

agentul meu, John Brockman, care a insistat până am acceptat să scriu această carte,

Bob Penna, Jesse Thaler și Jeremy England, pentru discuțiile lămuritoare despre quasari, sfaleroni și, respectiv, termodinamică cei de la care am primit reacții privind părți din manuscris, între care Mama, fratele meu Peter, Luisa Bahet, Rob Bensinger, Katerina Bergström, Erik Brynjolfsson, Daniela Chita, David Chalmers, Nima Deghani, Henry Lin, Elin Malmsköld, Toby

Ord, Jeremy Owen, Lucas Perry, Anthony Romero, Nate Soares și Jaan Tallinn,

supereroii care au comentat versiuni integrale ale întregii cărți, Meia, Tata, Anthony Aguirre, Paul Almond, Matthew Graves, Phillip Helbig, Richard Mallah, David Marble, Howard Messig, Luiño Seoane, Marin Solačić, editorul meu Dan Frank, iar mai presus de toate

Meia, iubita mea muză și tovarăș de călătorie, pentru încurajarea, sprijinul și inspirația pe care mi le-a oferit neîncetat și fără de care această carte n-ar fi putut exista.

VIAȚA 3.0

Crezi că IA supraumană ar putea fi creată în acest secol?

Nu

Da

Continuă
cu pagina
următoare

Sari la capitolul 1
(p. 31)

PRELUDIU

Povestea Echipei Omega

Echipa Omega era sufletul companiei. Dacă restul întreprinderii aducea bani pentru a face ca lucrurile să funcționeze, prin diverse aplicații comerciale ale inteligenței artificiale în sens restrâns, Echipa Omega avansa în căutarea a ceea ce directorul general al companiei visase dintotdeauna: crearea inteligenței artificiale generale. Majoritatea celorlalți angajați îi priveau pe „omegi”, așa cum îi numeau ei afectuos, ca pe niște visători aflați mereu la decenii distanță de scopul lor. Îi simpatizau totuși, fiindcă se bucurau de prestigiul pe care munca de pionierat a omegilor îl aducea companiei și foloseau algoritmi îmbunătățiți pe care omegii li-i puneau la dispoziție din când în când.

Ceea ce nu știau ei era că omegii își făuriseră această imagine ca să ascundă un secret: se aflau extrem de aproape de înfăptuirea celui mai îndrăzneț plan din istoria omenirii. Carismaticul director general îi alesese nu numai pentru că erau cercetători străluciți, ci și pentru ambiția, idealismul și dorința lor de a sluji omenirea. Le reamintea că planul lor era foarte primejdios, iar dacă guverne puternice l-ar afla, ar fi în stare de orice – inclusiv să-i răpească – pentru a-i opri sau, mai curând, pentru a le fura codul. Ei erau însă dedicați cu trup și suflet, cam din același motiv pentru care mulți mari fizicieni ai lumii se alăturaseră Proiectului Manhattan de

construire a bombei atomice: erau convinși că dacă n-o fac ei primii, o va face cineva mai puțin idealist.

Inteligența artificială creată de ei, numită Prometeu, se dezvoltă continuu. Deși capacitățile ei cognitive erau încă departe de cele ale oamenilor în multe domenii – în privința abilităților sociale, de pildă – omegii se străduiseră s-o perfecționeze pentru o sarcină precisă: programarea sistemelor de inteligență artificială (IA). Aleseseră anume această strategie, urmând ideea matematicianului britanic Irving Good enunțată încă din 1965: „Definim o mașină ultrainteligentă ca pe o mașină care poate întrece în toate activitățile intelectuale orice om, oricât de deștept ar fi. Din moment ce proiectarea mașinilor este una dintre aceste activități intelectuale, o mașină ultrainteligentă ar putea proiecta mașini încă și mai bune; ar avea loc atunci fără îndoială o «explozie a inteligenței», iar inteligența umană ar fi lăsată mult în urmă. Așadar, prima mașină ultrainteligentă ar fi ultima invenție pe care omul ar avea nevoie s-o facă, cu condiția ca mașina să fie suficient de docilă pentru a ne spune cum s-o ținem sub control.”

Ei s-au gândit că, dacă ar putea iniția această autoperfecționare recursivă, curând mașina ar deveni suficient de deșteaptă încât să deprindă singură toate celelalte abilități umane care i-ar fi folositoare.

PRIMELE MILIOANE

Era ora nouă dimineața, într-o zi de vineri, când s-au hotărât să-i dea drumul. Prometeu zumzăia în componentele lui electronice așezate pe șiruri lungi într-o încăpere cu acces limitat și aer condiționat. Din motive de siguranță era complet decuplat de internet, dar conținea o copie locală a unei mari părți a rețelei (Wikipedia, Biblioteca Congresului, Twitter, o selecție de pe Youtube, mare parte din Facebook etc.) pentru a-i folosi datele în procesul de învățare.* Au ales acest

* Pentru simplitate, am presupus în povestea mea că avem de-a face cu aceeași economie și tehnologie de azi, deși majoritatea cercetătorilor cred că de IA generală la nivel uman ne despart cel puțin decenii. Planul Omega ar

moment de lansare ca să lucreze netulburați: rudele și prietenii lor credeau că se află într-o casă de odihnă a companiei. Bucătăria era plină cu mâncare semipreparată și băuturi energizante, iar ei erau pregătiți să înceapă.

Când l-au lansat, Prometeu se descurca ceva mai slab decât ei la programarea sistemelor IA, dar compensa prin viteza uimitoare, făcând munca a o mie de oameni vreme de un an cât timp ei își beau paharul de Red Bull. Pe la 10 a.m. își încheiase prima reproiectare a lui însuși, v2.0, o versiune ceva mai bună, dar încă subumană. Când s-a lansat Prometeu 5.0 la 2 p.m., omegii au fost însă uluiți: le depășise cu mult performanța, iar ritmul progresului părea să se accelereze. La căderea serii s-au hotărât să-l angajeze pe Prometeu 10.0 în faza a doua a planului lor: să câștige bani.

Prima lor țință a fost MTurk, Amazon Mechanical Turk. După lansarea lui în 2005 ca piață de externalizare prin internet, a crescut rapid, zeci de mii de oameni din lumea întreagă concurând anonim douăzeci și patru de ore pe zi pentru a îndeplini activități structurate numite HTTs, „Human Intelligence Tasks” (sarcini care presupun inteligență umană). Spectrul acestor sarcini se întindea de la transcrierea înregistrărilor audio până la clasificarea imaginilor și redactarea descrierilor de pagini web, iar toate aveau ceva în comun: dacă reușeai să le îndeplinești, nimeni nu și-ar fi dat seama că ești o IA. Prometeu 10.0 a îndeplinit în mod acceptabil cam jumătate din categoriile de sarcini. Pentru fiecare asemenea categorie de sarcini, omegii l-au pus pe Prometeu să proiecteze un modul IA limitat, destinat numai acestor sarcini. Au încărcat apoi modulul pe Amazon Web Services, o platformă de *cloud-computing** pe care puteau rula atâtea mașini virtuale câte erau închiriate. La fiecare dolar plătit diviziei de *cloud-computing* de la Amazon, câștigau mai mult de doi dolari de la divizia MTurk

fi și mai ușor de realizat în viitor dacă economia digitală și-ar continua creșterea și ar fi disponibile online și mai multe servicii, fără a se cere informații în prealabil. (N. a.)

* *Cloud-computing* („calculare în nor”) presupune utilizarea puterii de calcul sau de stocare a serverelor aflate la distanță prin intermediul internetului, acele servere fiind închiriate la cerere. (N. t.)

a aceluiași Amazon. Amazon habar n-avea că exista asemenea posibilitate de speculă chiar în interiorul propriei companii!

Pentru a-și ascunde urmele, omegii creaseră discret în lunile precedente mii de conturi MTurk în numele unor persoane fictive, iar modulele Prometeu își asumau acum identitățile acestora. Clienții MTurk erau plătiți de regulă după vreo opt ore, moment în care omegii reinvesteau banii în mai mult timp de *cloud-computing*, folosind module ameliorate create de ultima versiune a lui Prometeu, aflat în continuă perfecționare. Pentru că își puteau dubla banii la fiecare opt ore, curând au saturat fondurile MTurk, și au constatat că nu pot câștiga mai mult de un milion de dolari pe zi fără să atragă atenția asupra lor. Dar asta era mai mult decât suficient pentru a-și finanța următorul pas, eliminând orice nevoie de a cere bani directorului financiar-contabil.

JOCURI PERICULOASE

Dincolo de descoperirile lor în IA, omegii se amuzau să-și imagineze cum pot câștiga bani cât mai rapid, odată lansat Prometeu. În fond, întreaga economie digitală le stătea la dispoziție, dar nu cumva era mai bine să înceapă să creeze jocuri pe calculator, muzică, filme sau software, să scrie cărți sau articole, să tranzacționeze acțiuni la bursă sau să facă invenții și să le vândă? Asta ar însemna să maximizeze rata de recuperare a investiției, dar strategiile normale de investiție erau o parodie derulată cu încetinitorul a ceea ce puteau face ei: în vreme ce un investitor normal ar fi mulțumit cu un beneficiu de 9% pe an, investițiile lor MTurk produceau 9% pe oră, generând de opt ori mai mulți bani în fiecare zi. Așa încât, după saturarea MTurk, ce urma?

S-au gândit mai întâi să distrugă bursa – la urma urmei, mulți dintre ei refuzaseră un job bine plătit în dezvoltarea IA pentru fonduri speculative, care investeau sume mari exact în aceeași idee. Unii își aminteau că în felul ăsta câștigase IA primele milioane în filmul *Transcendence*. Dar noile reglementări privind

instrumentele financiare derivate, după criza din anul precedent, le limitau alegerile. Au înțeles curând că, deși puteau avea beneficii mult mai mari decât alți investitori, era puțin probabil să obțină câștiguri comparabile cu ce le-ar aduce vânzarea propriilor produse. Când ai la dispoziție prima mașinărie superinteligentă din lume, e de preferat să investești în propriile tale companii decât în companiile altora! Deși puteau apărea unele excepții (cum ar fi folosirea talentelor supraumane de hacker ale lui Prometeu pentru a căpăta informații din interior și a cumpăra acțiuni care urmau să crească), omegii și-au spus că nu merita să atragă atenția asupra lor.

Când s-au concentrat asupra produselor pe care le puteau dezvolta și vinde, jocurile pe calculator li s-au părut cea mai bună alegere. Prometeu se putea specializa rapid în proiectarea unor jocuri atrăgătoare, manevrând cu ușurință codificarea, grafica, formarea imaginilor și toate celelalte sarcini necesare pentru a obține un produs finit. Mai mult, după digerarea tuturor datelor din rețea despre preferințele oamenilor, ar ști exact ce place fiecărei categorii de gameri, și ar dobândi o abilitate supraumană de a optimiza un joc pentru câștigul maxim. *The Elder Scrolls V: Skyrim*, un joc cu care omegii au pierdut mai mult timp decât erau dispuși să recunoască, adusesese câștiguri de peste 400 de milioane de dolari în prima săptămână de la lansare, în 2011, iar ei erau convinși că Prometeu putea crea ceva cel puțin la fel de tentant în douăzeci și patru de ore, folosind resurse de *cloud-computing* costând un milion de dolari. L-ar putea vinde apoi online și l-ar pune pe Prometeu să creeze personaje umane care vorbesc despre joc în blogosferă. Dacă asta le-ar aduce 250 de milioane de dolari într-o săptămână, ei și-ar putea dubla investiția de opt ori în opt zile, cu un beneficiu de 3% pe oră – mai mic decât cel de la început, cu MTurk, dar mult mai sigur. Dezvoltând o serie de alte jocuri în fiecare zi, și-au făcut socoteala că ar putea curând câștiga 10 miliarde de dolari, fără a satura piața jocurilor.

Dar o specialistă în cybersecuritate din echipa lor a analizat acest plan al jocurilor. Ea a arătat că ar apărea un risc inacceptabil dacă Prometeu ar evada și ar căpăta control asupra propriului

destin. Fiindcă nu puteau ști cum ar evolua obiectivele lui în decursul autoperfecționării recursive, au hotărât să fie prevăzători și să-l mențină „închis” pe Prometeu, ca să nu evadeze pe internet. Pentru mașinăria principală care rula în încăperea cu acces limitat, ei foloseau izolarea fizică: nu exista conexiune internet, iar singurul output de la Prometeu era sub forma unor mesaje și documente pe care le trimitea unui calculator controlat de omegi.

Însă pe un calculator conectat la internet, rularea oricărui program complicat creat de Prometeu era o propunere riscantă: cum omegii nu-și dădeau seama până la capăt ce-ar face, nu puteau ști dacă nu cumva ar începe, de pildă, să se răspândească viral online. Când au testat software-ul scris de Prometeu pentru sarcinile MTurk, s-au protejat de acest risc rulându-l doar în interiorul unei mașini virtuale. Acesta e un program care simulează un calculator: de exemplu, mulți utilizatori de Mac cumpără un software mașină virtuală care le permite să ruleze programele Windows făcându-le să creadă că e o mașină Windows veritabilă. Omegii își creaseră propria lor mașină virtuală, botezată Cutia Pandorei, care simula o mașină ultrasimplificată, lipsită de toate accesoriile pe care le asociem de regulă cu calculatoarele: nici un fel de tastatură, monitor, difuzoare, legătură internet – nimic. Pentru transcrierile audio MTurk, omegii au aranjat așa încât tot ce putea intra în Cutia Pandorei să fie un singur fișier audio, și tot ce putea ieși să fie un singur document text – transcrierea. Aceste „legi ale cutiei” erau pentru software-ul dinăuntru cam ce sunt legile fizicii pentru noi, în interiorul universului nostru: software-ul nu putea ieși din cutie, la fel cum noi nu putem călători cu viteze mai mari decât viteza luminii, oricât de deștepți am fi. Cu excepția acelor unice input și output, software-ul dinăuntru Cutiei Pandorei era efectiv prizonier într-un univers paralel, cu propriile lui reguli computaționale. Omegii se temeau atât de mult de evadare, încât au adăugat și o restricție temporală, limitând timpul de viață al codurilor în care n-aveau încredere. De pildă, de fiecare dată când software-ul de transcriere termina de transcris un fișier audio, întreaga memorie a Cutiei Pandorei era automat ștearsă, iar programul era reinstalat de la zero. Astfel, când

începea o nouă sarcină de transcriere, nu știa ce se întâmplase înainte, deci nu putea învăța în cursul timpului.

Când omegii au folosit norul Amazon pentru proiectul lor MTurk, ei au putut pune toate modulele create de Prometeu pentru îndeplinirea sarcinilor în asemenea cutii virtuale din nor, pentru că input-urile și output-urile MTurk erau foarte simple. Metoda nu mai funcționa însă la jocurile pe calculator cu multă grafică, care nu puteau fi „închise în cutie” fiindcă trebuiau să aibă acces deplin la întregul hardware al calculatorului jucătorului. Mai mult, ei n-au vrut să se expună riscului ca vreun utilizator expert să le analizeze codul jocului, să descopere Cutia Pandorei și să vrea să vadă ce-i înăuntrul ei. Riscul evadării scotea din calcul nu doar piața jocurilor, dar și piața extrem de profitabilă a altor tipuri de software, cu sute de miliarde de dolari care puteau fi câștigați.

PRIMELE MILIARDE

Omegii se concentraseră asupra produselor foarte scumpe, pur digitale (evitând o fabricare lentă) și ușor de înțeles (de pildă, texte sau filme despre care știau că nu prezintă riscul evadării). În cele din urmă se hotărâseră să lanseze o companie media și să înceapă cu desene animate. Website-ul, planul de marketing și comunicatele de presă fuseseră deja gata înainte ca Prometeu să devină superinteligent – nu lipsea decât conținutul.

Deși duminică dimineața capacitățile lui Prometeu deveniseră uimitoare și se extrăgeau constant bani de la MTurk, abilitățile lui intelectuale rămăseseră limitate: Prometeu fusese în mod deliberat optimizat pentru a proiecta sisteme IA și a scrie programe care executau sarcini plictisitoare pentru MTurk. De exemplu, nu se pricepea să facă filme, iar asta nu din vreun motiv profund, ci din același motiv pentru care James Cameron nu se pricepea să facă filme atunci când s-a născut: este o abilitate pe care o deprinzi în timp. La fel ca un copil uman, Prometeu putea învăța orice voia din datele la care avea acces. Dacă lui Cameron i-a luat ani de zile

ca să învețe să scrie și citească, Prometeu a reușit asta de vineri, când a avut timp și să citească întreaga Wikipedie și câteva milioane de cărți. Să faci filme era însă mai greu. A scrie un scenariu pe care oamenii să-l găsească interesant era la fel de greu ca a scrie o carte, cerând înțelegerea detaliată a societății umane și a ceea ce îi atrage pe oameni. Transformarea scenariului într-un fișier video presupunea o cantitate enormă de calcule pentru formarea imaginilor cu actorii simulați și cu decorul complex prin care se mișcau, voci simulate, o coloană sonoră potrivită etc. Duminică dimineața Prometeu a urmărit un film de două ore într-un minut, timp în care a citit și cărțile pe care se baza filmul și toate cronicile și evaluările online. Omegii au observat că, după ce Prometeu a vizionat câteva sute de filme, a început să prevadă corect ce gen de cronici s-ar putea scrie despre un film și cum ar fi apreciat de diferite tipuri de public. A învățat să scrie cronici pătrunzătoare, comentând totul, de la intrigă și jocul actorilor la detalii tehnice cum ar fi lumina și unghiul de filmare. Omegii au dedus de aici că, atunci când își va face propriile filme, Prometeu va ști ce înseamnă succesul.

Omegii l-au învățat pe Prometeu să înceapă cu animația, fără să se întrebe cine sunt actorii simulați. Duminică noaptea și-au încheiat furtunosul weekend înarmându-se cu bere și popcorn, stingând luminile și urmărind filmul de debut al lui Prometeu. Era o comedie-fantezie animată, în genul lui *Frozen*, iar formarea imaginilor fusese realizată cu un cod „închis” al lui Prometeu rulat în norul Amazon, cheltuindu-se astfel aproape tot milionul de dolari câștigat în acea zi de la MTurk. Când a început filmul, omegilor li s-a părut deopotrivă fascinant și înspăimântător faptul că era creația unei mașini care nu fusese îndrumată de om. Curând însă, râdeau la poante și își țineau respirația în momentele dramatice. Unii chiar au lăcrimat la sfârșitul emoționant, atât de cufundați în această realitate ficțională, încât au uitat cine e autorul.

Omegii au plănuțit ca lansarea site-ului să aibă loc vineri, lăsând timp lui Prometeu să producă mai mult conținut și lor să se ocupe de alte lucruri: să cumpere publicitate și să recruteze angajați pentru companiile fictive pe care le înființaseră în ultimele luni.

Pentru a-și ascunde urmele, povestea oficială era că propria lor companie media (fără vreo legătură directă cu ei) a cumpărat cea mai mare parte a conținutului de la producători de film independenți, de regulă startup-uri de înaltă tehnologie din zone sărace. Acești falși furnizori au fost plasați în locuri îndepărtate, cum ar fi Tiruchchirappalli și Idukki, pe care nici cei mai curioși jurnaliști nu s-ar osteni să le viziteze. Singurii angajați pe care i-au recrutat efectiv lucrau în marketing și administrație, și oricui ar fi pus întrebări le-ar fi răspuns că echipa lor de producție se afla în altă parte, iar deocamdată nu dădea interviuri. Pentru a se potrivi cu povestea oficială, au ales drept slogan al companiei „Să canalizăm talentul creator al lumii”, și au ales ca brand ideea că sunt radical diferiți prin folosirea tehnologiei de vârf pentru a-i ajuta pe oamenii creativi, în special pe cei din țările în curs de dezvoltare.

Vineri, vizitatorii curioși care au intrat pe site-ul lor au dat peste ceva ce amintea de serviciile Netflix și Hulu, dar cu o diferență semnificativă: toate seriile animate erau noi. Păreau captivante – majoritatea seriilor constau în episoade de 45 de minute, cu un scenariu bine articulat, fiecare încheindu-se într-un fel care te făcea nerăbdător să afli ce se întâmplă în episodul următor. Și erau mai ieftine decât cele produse de concurență. Primul episod al fiecărui serial era gratuit, și le puteai vedea pe celelalte plătiind 49 de cenți pe episod, cu reduceri pentru întregul serial. La început existau doar trei seriale cu trei episoade fiecare, dar zilnic erau adăugate noi episoade, satisfăcând diverse gusturi. În primele două săptămâni ale lui Prometeu, talentele lui de cineast au sporit rapid, nu numai în privința calității filmelor, dar și a algoritmilor pentru simularea personajelor și formarea imaginilor, ceea ce a redus mult costul pentru *cloud-computing* la fiecare nou episod. Ca atare, omegii au putut lansa zeci de noi seriale în prima lună, publicul-țintă cuprinzând toate categoriile, de la copii la adulți, și s-au extins pe toate piețele importante din lume corespunzând limbilor de largă circulație, ceea ce a dat site-ului un caracter internațional incomparabil cu al concurenților. Unii comentatori au fost impresionați de faptul că nu doar banda sonoră era în mai multe limbi, dar și imaginile înseși: de pildă, când un personaj vorbea

în italiană, mișcarea buzelor se potrivea cuvintelor italienești, la fel și gestică tipic italiană. Deși Prometeu putea crea acum filme cu actori simulați imposibil de deosebit de oamenii reali, omegii n-au făcut-o, pentru ca nu cumva să se dea de gol. Au lansat însă multe seriale cu personaje umane animate semirealiste, în genul show-urilor TV și filmelor tradiționale.

Rețeaua lor s-a bucurat de mare succes, iar numărul vizionărilor a crescut spectaculos. Mulți fani au găsit că personajele și scenariile sunt mai inteligente și mai interesante decât în cele mai scumpe producții hollywoodiene pentru marele ecran, și erau încântați că le pot urmări la prețuri rezonabile. Beneficiind de o publicitate agresivă (omegii și-o permiteau fiindcă producția nu-i costa aproape nimic), de o acoperire media excelentă și cronici entuziaste, după o lună de la lansare venitul lor global a ajuns la zece milioane de dolari pe zi. Peste două luni depășiseră Netflix, iar peste trei au ajuns la o sută de milioane de dolari pe zi, începând să rivalizeze cu Time Warner, Disney, Comcast și Fox, ca unul dintre cele mai mari imperii media din lume.

Acest senzațional succes a atras atenția asupra lor, s-a spus că ar poseda o IA puternică, însă folosind doar o mică parte din venit, omegii au desfășurat o reușită campanie de dezinformare. Dintr-un sediu nou și kitschos din Manhattan, purtători de cuvânt proaspăt angajați relatau în amănunt poveștile lor inventate. Mulți oameni au fost angajați în chip de diversiune, între care scenariști adevărați din lumea întreagă, pentru a elabora noi seriale, fără ca vreunul să aibă habar de Prometeu. Derutanta rețea internațională de subcontractori a făcut cu puțință ca majoritatea angajaților să presupună că alții de altundeva fac cea mai mare parte din muncă.

Pentru a deveni mai puțin vulnerabili și a evita bănuielile provocate de excesul de *cloud computing*, au angajat de asemenea ingineri ca să construiască o serie de centre puternice de calculatoare pretutindeni în lume, deținute de companii fictive aparent neafiliate. Deși numite „centre verzi de date”, fiindcă erau alimentate în mare parte cu energie solară, acestea se concentrau de fapt asupra calculului, nu asupra stocării. Prometeu le proiectase până în cele mai mici detalii, folosind doar hardware-ul disponibil pe piață și

optimizându-le pentru a minimiza timpul de construcție. Cei care le-au construit și le conduceau habar n-aveau ce se calcula acolo, își închipuiau că au de-a face cu instalații comerciale de *cloud-computing*, asemănătoare celor deținute de Amazon, Google și Microsoft, și știau doar că toate vânzările sunt gestionate de undeva de departe.

NOI TEHNOLOGII

Pe măsură ce treceau lunile, imperiul de afaceri condus de omegii pătrundea în tot mai multe domenii ale economiei mondiale, grație planificării supraumane a lui Prometeu. Analizând atent datele internaționale, încă din prima săptămână le oferise omegilor un plan detaliat de creștere pas cu pas, plan continuu îmbunătățit odată cu acumularea datelor și resurselor de calcul. Deși Prometeu era departe de a fi atotcunoscător, capacitățile lui depășiseră atât de mult limitele omenescului, încât omegii îl priveau ca pe oracolul perfect, oferind supus răspunsuri și sfaturi strălucite la toate întrebările lor.

Software-ul lui Prometeu era acum optimizat pentru a obține maximumul din hardul mai curând mediocru, inventat de oameni, pe care rula, și, după cum anticipaseră omegii, Prometeu a găsit căi de a îmbunătăți spectaculos acest hardware. Temându-se de o evadare, ei au refuzat construcția robotică a unor dispozitive pe care Prometeu ar fi controlat-o direct. Au angajat în schimb numeroși oameni de știință și ingineri de înaltă clasă din diferite colțuri ale lumii și le-au oferit rapoartele interne de cercetare scrise de Prometeu, pretinzând că erau ale unor cercetători din alte locuri. Aceste rapoarte prezentau în detaliu noi efecte fizice și tehnici de producție, pe care curând inginerii lor le-au testat, înțeles și stăpânit. Desigur, ciclurile umane normale de cercetare și dezvoltare durează ani, în mare măsură fiindcă ele presupun multe cicluri lente de încercări și eșecuri. Acum situația era diferită: Prometeu cunoștea deja următorii pași de urmat, așa încât erau limitați doar de viteza cu care oamenii puteau fi îndrumați să înțeleagă și să

construiască ce trebuia. Un profesor bun îi poate ajuta pe studenți să învețe știința mult mai repede decât dacă ei ar descoperi totul pe cont propriu pornind de la zero, iar Prometeu a făcut pe ascuns același lucru cu acești cercetători. Cum Prometeu putea prezice cu exactitate cât le-ar lua oamenilor să înțeleagă și să construiască dispozitive cu diferite instrumente, a găsit cea mai scurtă cale de înaintare, dând prioritate noilor instrumente care ar fi imediat înțelese și construite, și care ar ajuta la dezvoltarea unor instrumente și mai avansate.

Echipele de ingineri au fost astfel încurajate să-și folosească propriile mașini pentru a construi unele mai bune. Această autoaprovizionare nu numai că a adus economii bănești, dar i-a făcut pe omegi mai puțin vulnerabili la viitoare amenințări din lumea exterioară. În doi ani, produceau un hardware incomparabil mai bun decât existase vreodată în lume. Pentru a nu-și ajuta concurența, au păstrat secretă această tehnologie, folosind-o doar pentru a-l upgrada pe Prometeu.

Ce a observat totuși lumea a fost o uimitoare explozie tehnologică. Companii nou înființate de pretutindeni lansau noi produse revoluționare aproape în toate domeniile. Un startup sud-coreean a lansat un nou acumulator care stoca de două ori mai multă energie decât acumulatorul de la laptopul tău, fiind de două ori mai ușor, și se încărca în mai puțin de un minut. O firmă finlandeză a venit cu un panou solar ieftin de două ori mai eficient decât al concurenților. O companie germană a anunțat un nou tip de cablu care putea fi produs pe scară largă și era supraconductor la temperatura camerei, revoluționând sectorul energetic. Un grup de cercetare în biotehnologie cu sediul la Boston a anunțat că se afla în faza a doua a testelor clinice primul medicament eficace pentru slăbit, fără efecte secundare, și se zvonea că un producător indian vindea deja așa ceva pe piața neagră. O companie californiană a ripostat anunțând faza a doua a testelor pentru un medicament-minune anticancer, care ajută sistemul imunitar să recunoască și să atace celulele care prezintă unele dintre cele mai frecvente mutații cancerigene. Exemplele veneau în cascadă, întărind ideea că era vorba de o nouă epocă de aur a științei. Pe de altă parte,

companii de robotică răsăreau pretutindeni în lume ca ciupercile. Nici unul dintre roboți nu se apropia de inteligența umană, iar cei mai mulți dintre ei nu semănau deloc cu oamenii. Ei au perturbat însă violent economia, înlocuindu-i treptat pe cei mai mulți lucrători din industrie, transport, comerțul cu amănuntul și cu ridicata, construcții, minerit, agricultură, silvicultură și pescuit.

Ce nu s-a observat, grație eforturilor unei echipe de avocați experți, a fost că toate aceste firme erau controlate prin intermediari de omegă. Prometeu invada birourile de brevete cu invenții senzaționale prin diferiți reprezentanți, iar aceste invenții au condus treptat la dominarea tuturor sectoarelor tehnologiei.

Deși noile companii și-au făcut inamici puternici în rândul concurenței, ele și-au făcut încă și mai mulți prieteni puternici. Erau extraordinar de profitabile, și sub sloganuri de genul „Investim în comunitatea noastră” cheltuiau părți însemnate din profit angajând oameni pentru proiecte comunitare – adesea aceiași oameni care fuseseră concediați de companiile rivale aduse în pragul falimentului. Omegii foloseau analize ale lui Prometeu ca să identifice joburi optime pentru angajați și comunitate, cu costuri minime, adaptate condițiilor locale. În regiunile cu un nivel de trai ridicat se concentrau adesea asupra întării comunității, familiei și culturii, iar în regiunile sărace se preocupau în plus de înființarea și întreținerea școlilor, sănătate, ajutorarea persoanelor în vârstă, locuințe la prețuri modice, spații verzi și infrastructura de bază. Mai peste tot cetățenii erau de părere că asemenea lucruri trebuiau făcute de multă vreme. Politicienii locali primeau donații generoase și încurajau aceste investiții comunitare.

DOBÂNDIREA PUTERII

Omegii creaseră o companie media nu doar pentru a-și finanța primele întreprinderi tehnologice, dar și în vederea următorului pas al ambițiosului lor plan: stăpânirea lumii. La un an după lansare, au înființat pretutindeni în lume canale de știri remarcabil

de bune. Spre deosebire de celelalte canale, ele au fost în mod deliberat concepute pentru a merge în pierdere și au fost considerate un serviciu public. De fapt, canalele lor de știri nu produceau nici un fel de venit: nu dădeau publicitate și erau accesibile gratuit oricui avea conexiune internet. Restul imperiului lor media aducea atâtea câștiguri, încât își permiteau să cheltuiască pe serviciul de știri mult mai mult decât orice întreprindere jurnalistică din istorie – iar roadele s-au văzut. Prin recrutare agresivă, oferind salarii extrem de competitive ziariștilor și reporterilor de investigație, au făcut dezvăluiri senzaționale cu un remarcabil talent. Folosind serviciile unei rețele globale care plătea pe oricine venea cu o știre interesantă, de la corupția locală până la un eveniment emoționant, ei erau de regulă primii care o anunțau. Cel puțin asta credea lumea: de fapt, adesea ei erau primii fiindcă relatările atribuite jurnaliștilor fuseseră descoperite de Prometeu prin monitorizarea în timp real a internetului. Pe toate aceste site-uri video de știri apăreau de asemenea articole.

Prima fază a strategiei legate de știri presupunea câștigarea încrederii oamenilor, ceea ce au reușit cu mare succes. Acceptând să piardă bani, au obținut o remarcabilă acoperire locală și regională, iar ziariștii de investigație dezvăluiau deseori scandaluri care țineau trează atenția publicului. Dacă într-o țară divizarea politică era accentuată și lumea se obișnuise cu știri partizane, ei lansau câte un canal de știri pentru fiecare facțiune, deținut chipurile mereu de câte o altă companie, câștigând treptat încrederea acelei facțiuni. Unde era posibil, apelau la intermediari pentru a cumpăra cele mai influente canale existente, iar apoi le îmbunătățeau eliminând publicitatea și introducând propriul lor conținut. În țările în care cenzura și ingerințele politicului le amenințau eforturile, acceptau la început orice le cerea guvernul pentru a rezista, cu sloganul intern secret „Adevărul și numai adevărul, însă poate nu întreg adevărul”. De regulă, Prometeu oferea sfaturi excelente în asemenea situații, stabilind ce politicieni trebuie prezentați favorabil și ce politicieni (în general corupții locali) pot fi expuși. Prometeu le spunea de asemenea ce sfori trebuie trase, cine și cum trebuie mituit.

Această strategie a avut un succes zdrobitor pretutindeni în lume, canalele controlate de omegi devenind cele mai de încredere surse de știri. Chiar și în țările unde guvernele încercau să contracareze creșterea popularității lor și-au dobândit credibilitate, iar multe din știrile lor se răspândeau pe căi neoficiale. Directorii canalelor de știri concurente simțeau că bătălia era pierdută: cum poți să obții profit în competiție cu cineva mai bine finanțat, care își oferă produsele gratis? Cu audiența în scădere, tot mai multe rețele hotărau să-și vândă canalele de știri – de regulă vreunui consorțiu care se dovedea apoi că e controlat de omegi.

La vreo doi ani de la lansarea lui Prometeu, când faza de câștigare a încrederii era în mare parte încheiată, omegii au inițiat faza a doua a strategiei lor legate de știri: persuasiunea. Încă dinainte observatorii fini au sesizat indicii ale agendei politice aflate în spatele noilor medii de informare: părea să existe o ușoară tendință spre centru, departe de extremismele de tot felul. Mulțimea lor de canale orientate către diferite grupuri reflecta în continuare animozitatea dintre Statele Unite și Rusia, India și Pakistan, diverse religii, facțiuni politice etc., dar tonul critic era ușor îndulcit, în prim-plan aflându-se teme concrete implicând banii și puterea, iar nu atacurile la persoană, răspândirea panicii și zvonurile fără acoperire. Odată începută cu adevărat faza a doua, această tendință de dezamorsare a vechilor conflicte a devenit tot mai vizibilă, prezentându-se frecvent relatări despre înțelegerile dintre adversarii tradiționali, împreună cu investigații care arătau că mulți susținători vocali ai conflictelor erau mânați de interese materiale personale.

Comentatorii politici au observat că, în paralel cu atenuarea conflictelor regionale, părea să apară o presiune concertată către reducerea amenințărilor globale. De pildă, riscurile unui război nuclear au început brusc să fie discutate pretutindeni. Câteva filme de mare succes se bazau pe scenarii în care un război nuclear era declanșat din întâmplare sau în mod deliberat, prezentând urmările lui distopice – iarna nucleară, prăbușirea infrastructurii, foametea generalizată. Documentare abil concepute arătau ce impact ar avea iarna nucleară asupra fiecărei țări. Oamenilor de

știință și politicienilor care susțineau dezangajarea nucleară li se acorda un timp amplu de emisie, în special pentru a discuta rezultatele unor noi studii privind măsurile ce pot fi luate – studii finanțate de organizații științifice care primiseră donații generoase de la noile companii de tehnologie. Ca atare, a apărut tendința politică de a se depăși starea de alertă și a reduce arsenalele nucleare. Mediile de informare acordau de asemenea o atenție sporită modificării globale a climei, subliniind adesea inovațiile tehnologice datorate lui Prometeu, care scăzuseră mult prețul energiei regenerabile, și încurajând guvernele să investească în asemenea noi infrastructuri energetice.

În paralel cu preluarea controlului asupra mediilor de informare, omegii l-au folosit pe Prometeu pentru a revoluționa educația. Pornind de la cunoștințele și înzestrările fiecăruia în parte, Prometeu putea găsi calea cea mai rapidă de învățare a unei materii noi, așa încât fiecare să fie atras și motivat să persevereze, și putea produce materiale video, texte, exerciții și alte instrumente didactice optimizate. Companiile controlate de omegi au lansat deci pe piață cursuri online despre aproape orice, adaptate nu doar mediului lingvistic și cultural, dar și nivelului de pornire al elevului. Indiferent dacă erai un analfabet de 43 de ani care voia să învețe să citească sau un doctor în biologie interesat de ultimele noutăți în imunoterapia cancerului, Prometeu avea pentru tine cursul ideal. Aceste oferte nu semănau deloc cu celelalte cursuri online: profitând de talentele de cineast ale lui Prometeu, segmentele video te incitau cu adevărat prin metafore sugestive cu efect direct asupra ta, făcându-te să-ți dorești să afli mai multe. Unele cursuri erau contra cost, dar multe erau gratuite, spre încântarea profesorilor din lumea întreagă care le puteau folosi la școală – și a oricui era doritor să învețe ceva.

Aceste supraputeri educaționale s-au dovedit a fi instrumente eficiente în scopurile politice, creând „seriale de persuasiune”, în care dezvăluirile dintr-un film te făceau să-ți revizuiști opiniile și te îndemnau să privești un alt film despre un subiect înrudit, pesemne la fel de convingător. Când scopul era, de pildă, dezamorsarea unui conflict între două națiuni, în ambele țări erau

difuzate independent documentare istorice care prezentau originea și evoluția conflictului într-o lumină mai nuanțată. Știri cu caracter pedagogic îi demascau pe aceia din propria tabără care trăgeau beneficii din perpetuarea conflictului și prezentau tehnicile prin care ei îl alimentau. În același timp, figuri agreabile din cealaltă națiune începeau să apară în emisiuni populare ale canalelor de divertisment, la fel cum în trecut personaje prezentate într-o lumină favorabilă au susținut mișcările de emancipare civică și cele pentru drepturile persoanelor gay.

Curând, comentatorii politici au observat creșterea sprijinului în favoarea unei agende concentrate în jurul a șapte sloganuri:

1. Democrație
2. Reduceri fiscale
3. Reduceri ale serviciilor sociale guvernamentale
4. Reducerea cheltuielilor militare
5. Comerț liber
6. Granițe deschise
7. Companii responsabile social

Mai puțin evident era scopul subteran: erodarea tuturor structurilor de putere din lume. Punctele 2-7 erodau puterea statului, iar democratizarea lumii dădea imperiului de afaceri al omegilor o mai mare influență în selectarea liderilor politici. Companiile responsabile social slăbeau puterea statului preluând tot mai multe servicii pe care guvernele le ofereau (sau ar fi trebuit să le ofere). Elita de afaceri tradițională era slăbită pur și simplu pentru că nu putea concura pe piața liberă cu companiile susținute de Prometeu, așa încât deținea o parte tot mai mică din economia mondială. Liderilor de opinie tradiționali, de la partide politice la grupări religioase, le lipsea mecanismul de persuasiune pentru a face față imperiului media al omegilor.

Așa cum se întâmplă în orice schimbare radicală, existau învingători și învinși. Deși în majoritatea țărilor se simțea o undă de optimism, odată ce educația, serviciile sociale și infrastructura se îmbunătățiseră, conflictele se domoliseră, iar companiile lansau tehnologii inovatoare care schimbau viețile oamenilor, nu toată

lumea era fericită. În vreme ce mulți muncitori concediați erau reangajați pentru proiecte comunitare, cei care deținuseră putere și bogăție au văzut în general că ambele se topesc. Fenomenul a început în sectoarele media și tehnologic, dar s-a răspândit aproape pretutindeni. Diminuarea conflictelor a provocat reduceri ale bugetelor de apărare care i-au atins pe furnizorii militari. De regulă, companiile nou înființate nu erau cotate la bursă, cu justificarea că acționarii care vor să obțină profit maxim ar obstrucționa proiectele comunitare. Prin urmare, bursa s-a devalorizat la nivel mondial, amenințându-i deopotrivă pe magnații finanțelor și pe cetățenii de rând care contau pe fondurile lor de pensii. De parcă scăderea profiturilor companiilor listate la bursă n-ar fi fost de ajuns, firmele de investiții din lumea întreagă au observat o tendință îngrijorătoare: toți algoritmiile lor comerciali care în trecut avuseseră succes păreau să nu mai funcționeze, nereușind să se descurce nici măcar cu fondurile indexate. Cineva le bătea mereu pe propriul lor teren de joc.

Deși numeroși oameni puternici se opuneau valului schimbării, reacția lor era absolut inefficientă, de parcă ar fi căzut într-o capcană abil ticluită. Schimbări uriașe aveau loc într-un ritm atât de amețitor, încât era greu să le urmărești și să vii cu un răspuns coordonat. În plus, nu era deloc limpede ce trebuia făcut. Dreapta politică tradițională văzuse că-i sunt preluate majoritatea sloganelor, dar reducerea impozitelor și climatul de afaceri îmbunătățit ajutau mai curând concurența de înaltă tehnologie. Practic toate întreprinderile tradiționale cereau sprijin financiar, dar fondurile guvernamentale erau limitate, așa că se luptau fără speranță între ele, în vreme ce mediile de informare le prezentau ca pe niște dinozauri care cer subvenții de la stat pur și simplu din lipsă de competitivitate. Stânga politică tradițională se opunea liberului schimb și reducerii serviciilor sociale, dar era încântată de reducerea cheltuielilor militare și de scăderea sărăciei. Protestele ei erau amuțite de incontestabilul fapt că serviciile sociale se îmbunătățiseră odată ce erau oferite de companii idealiste, iar nu de stat. Sondajele de opinie arătau că majoritatea respondenților din lumea întreagă percepeau o îmbunătățire a calității vieții

și credeau că lucrurile se îndreaptă în general într-o direcție bună. Explicația de ordin matematic era simplă: înainte de Prometeu, jumătatea mai săracă a planetei deținea doar 4% din venitul global, ceea ce a permis companiilor controlate de omegi să câștige inițiile (și voturile) celor săraci, împărțind cu ei doar o mică parte din profituri.

CONSOLIDAREA

Ca urmare, partidele care îmbrățișau cele șapte sloganuri ale omeșilor au obținut pretutindeni victorii răsunătoare în alegeri. În campanii atent optimizate, ele se declarau de centru, denunțând lăcomia dreptei care cerea subvenții de la stat și căuta să declanșeze conflicte, și acuzând stânga că vrea să introducă mereu impozite. Nimeni n-a observat că Prometeu își alesese cu grijă candidații și trăgea sforile pentru a le asigura victoria.

Înainte de Prometeu, luase avânt o mișcare în favoarea unui venit de bază universal, care propunea un venit minim pentru toată lumea, obținut prin impozitare, ca remediu la șomajul tehnologic. Mișcarea a implodat atunci când au apărut proiectele comunitare ale corporațiilor, fiindcă imperiul de afaceri controlat de omegi oferea de fapt exact același lucru. Sub pretextul îmbunătățirii coordonării proiectelor lor comunitare, un grup internațional de companii a înființat Alianța Umanitară, o organizație nongvernamentală având drept scop identificarea și subvenționarea celor mai importante eforturi umanitare pretutindeni în lume. Curând, a ajuns s-o susțină practic întregul imperiu Omega, și s-au lansat proiecte globale la o scară fără precedent, chiar și în țări ocolite de explozia tehnologică, făcând să crească nivelul educației, asistenței medicale, prosperității și administrației. Evident, din culise Prometeu oferea planuri minuțios elaborate: în loc să împrăști bani cu țârăita, ca în propunerile pentru un venit de bază, Alianța (așa cum începuse să fie numită) îi privilegia pe cei care lucrau pentru cauza ei. Ca urmare, o mare parte din populația

lumii era recunoscătoare și fidelă Alianței – deseori mai mult decât față de propriul guvern.

Cu trecerea timpului, Alianța și-a asumat un rol tot mai important în conducerea lumii, iar puterea guvernelor naționale s-a erodat continuu. Bugetele naționale scădeau prin reducerea impozitelor, în vreme de bugetul Alianței depășea cu mult toate bugetele naționale însumate. Rolurile tradiționale ale guvernelor au devenit redundante și irelevante. Alianța oferea de departe cele mai bune servicii sociale, educație și infrastructură. Mediile de informare dezamorsaseră conflictele internaționale până în punctul în care cheltuielile militare au devenit aproape inutile, iar creșterea prosperității eliminase majoritatea cauzelor vechilor conflicte, bazate pe lupta pentru resurse insuficiente. Câțiva dictatori s-au opus violent acestei noi ordini mondiale și au refuzat să fie cum-părați, dar au fost cu toții răsturnați prin lovituri de stat și revolte atent orchestrate.

Omegii au încheiat cea mai spectaculoasă tranziție din istoria vieții pe Pământ. Planeta era pentru prima dată condusă de o unică putere, amplificată de o inteligență atât de vastă, încât putea permite vieții să prospere vreme de miliarde de ani pe Pământ și în cosmos – dar care erau de fapt planurile lor?

*

* *

Aceasta a fost povestea Echipei Omega. Restul cărții spune o altă poveste – una care n-a fost încă scrisă: povestea viitorului nostru cu IA. Cum ați vrea să se desfășoare? S-ar putea întâmpla ceva care să semene oarecum cu povestea Omega? Ați vrea să fie așa? Lăsând deoparte speculațiile privind IA supraumană, cum ați vrea să înceapă povestea noastră? Cum ați vrea să influențeze IA locurile de muncă, legile și armamentul în deceniul următor? Privind mai departe, cum ați scrie sfârșitul ei? Această poveste are într-adevăr proporții cosmice, odată ce implică nici mai mult, nici mai puțin decât soarta vieții în universul nostru. Și e o poveste pe care noi o vom scrie.

CAPITOLUL 1

Sunteți invitați la cea mai importantă dezbateră a epocii noastre

*Tehnologia oferă vieții posibilitatea de a prospera ca
niciodată până acum – sau de a se autodistruge.*

Institutul pentru Viitorul Vieții

La 13,8 miliarde de ani de la nașterea lui, universul nostru s-a trezit și devine conștient de sine. De pe o mică planetă albastră, minuscule părți conștiente ale universului au început să privească spre cosmos cu telescoape, descoperind în repetate rânduri că tot ce credeau că există e doar o infimă parte din ceva mult mai vast: un sistem solar, o galaxie și un univers cu peste o sută de miliarde de alte galaxii aranjate într-o structură complicată de grupuri, roiuri și superroiuri. Deși acești privitori la stele conștienți de sine se contrazic în multe privințe, ei cad de regulă de acord asupra faptului că galaxiile sunt frumoase și inspiră un sentiment înălțător.

Dar frumusețea este în ochii privitorului, nu în legile fizicii, așa încât, înainte ca universul nostru să se trezească, nu exista nici o frumusețe. Asta face și mai miraculoasă deșteptarea noastră cosmică – ea a transformat universul dintr-un zombi fără conștiință de sine într-un ecosistem viu adăpostind autorefecția, frumosul și speranța, căutarea unor obiective și a unei semnificații. Dacă nu s-ar fi deșteptat, universul ar fi rămas, cred eu, complet inutil, o uriașă irosire de spațiu. Dacă universul va adormi pe veci la loc, din pricina vreunei catastrofe cosmice sau a autodistrugerii, va deveni lipsit de sens.

Pe de altă parte, lucrurile ar putea merge și mai bine. Deocamdată nu știm dacă noi, oamenii, suntem singurii privitori la stele din univers, sau măcar primii, dar am aflat deja destule despre universul nostru pentru a ști că se poate deștepta încă mai mult decât a făcut-o până acum. Suntem pesemne la fel ca primul licăr al conștiinței de sine care apare atunci când te trezești dimineața din somn: premoniția unei conștiințe mult mai vaste la care ajungi după ce deschizi ochii și ești pe deplin treaz. Poate că viața se va răspândi prin cosmos și va prospera vreme de miliarde sau bilioane de ani – și poate că asta se va datora hotărârilor pe care le luăm aici, pe mica noastră planetă, în cursul vieții noastre.

O SCURTĂ ISTORIE A COMPLEXITĂȚII

Cum a apărut așadar această miraculoasă deșteptare? N-a fost un eveniment izolat, ci un pas dintr-un neîncetat proces durând 13,8 miliarde de ani, care face ca universul nostru să devină tot mai complex și mai interesant – și care continuă într-un ritm accelerat.

Ca fizician, mă consider norocos că am petrecut o bună parte din ultimul sfert de veac contribuind la înțelegerea istoriei noastre cosmice, și a fost o extraordinară aventură a cunoașterii. Pe când absolveam facultatea ne întrebam dacă vârsta universului este de 10 sau 20 de miliarde de ani, acum ne întrebăm dacă vârsta lui este de 13,7 sau 13,8 miliarde de ani, iar asta grație unor mai bune telescoape și calculatoare, și unei mai bune înțelegeri. Nu știm încă exact ce a declanșat big bang-ul sau dacă a fost într-adevăr începutul tuturor lucrurilor, ori doar urmarea unei etape anterioare. Înțelegem totuși destul de bine ce s-a întâmplat *după* big bang, grație unei avalanșe de măsurători de înaltă precizie, așa încât permițezi-mi să rezum cei 13,8 miliarde de ani ai istoriei cosmice.

La început a fost lumină. În prima fracțiune de secundă după big bang, întreaga parte a spațiului pe care o pot în principiu observa telescoapele noastre („universul observabil” sau „universul nostru”) era mult mai fierbinte și mai strălucitoare decât miezul

Soarelui nostru, și se extindea rapid. Deși sună spectaculos, era banal, fiindcă universul nostru nu conținea decât o supă densă, fierbinte, plictisitor de uniformă și lipsită de viață, alcătuită din particule elementare. Lucrurile arătau pretutindeni cam la fel, iar singura structură interesantă era dată de niște slabe unde acustice aleatorii care făceau ca supa să fie cu vreo 0,001% mai densă în anumite locuri. Aceste unde slabe se crede că au apărut ca fluctuații cuantice – pentru că principiul incertitudinii al lui Heisenberg din mecanica cuantică nu permite nici unui lucru să fie cu desăvârșire plictisitor și uniform.

Pe măsură ce universul nostru se extindea și se răcea, devenea tot mai interesant, odată ce particulele lui se combinau în obiecte mai complexe. În cursul primei fracțiuni de secundă, forța nucleară tare a adunat cuarcii în protoni (nuclee de hidrogen) și neutroni, iar în câteva minute unii dintre ei au fuzionat, formând nuclee de heliu. Peste vreo 400 000 de ani, forța electromagnetică a combinat aceste nucleele cu electroni pentru a forma primii atomi. Pe măsură ce universul continua să se extindă, atomii se răceau treptat, alcătuind un gaz întunecat și rece, iar întunericul acestei prime nopți a durat circa 100 de milioane de ani. După lunga noapte s-au ivit zorii cosmosului nostru, când forța gravitațională a reușit să amplifice acele fluctuații din gaz, atrăgând atomii laolaltă pentru a forma primele stele și galaxii. Aceste prime stele au produs căldură și lumină făcând să fuzioneze atomii de hidrogen în atomi mai grei cum sunt cei de carbon, oxigen și siliciu. Când primele stele au murit, mulți dintre atomii creați de ele au fost reciclați în cosmos, alcătuind planete ce se roteau în jurul stelelor din generația a doua.

La un moment dat, un grup de atomi s-a aranjat într-o structură complexă care se putea menține și replica. Curând, au apărut două exemplare identice, apoi numărul lor a continuat să se dubleze. Sunt de ajuns 40 de dublări pentru a se ajunge la un bilion, așa încât acest prim replicator a devenit repede o forță care trebuia luată în considerare. Apăruse viața.

CELE TREI ETAPE ALE VIEȚII

Problema definirii vieții e larg dezbătută. Există o mulțime de definiții, unele incluzând cerințe extrem de precise, cum ar fi alcătuirea din celule, care ar putea descalifica atât viitoarele mașini inteligente, cât și civilizațiile extraterestre. Fiindcă nu vrem să ne limităm perspectiva asupra viitorului vieții la speciile întâlnite până acum, vă propun să definim viața în sens foarte larg, ca pe un proces care își poate menține și replica propria-i complexitate. Nu materia (alcătuită din atomi) este replicată, ci informația (alcătuită din biți) care precizează cum sunt aranjați atomii. Când o bacterie face o copie a ADN-ului ei, nu sunt creați noi atomi, ci un nou set de atomi e aranjat în același tipar ca originalul, copiindu-se astfel informația. Cu alte cuvinte, putem privi viața ca pe un sistem de prelucrare a informației autoreplicator a cărui informație (software) îi determină deopotrivă comportamentul și planurile pentru hardware.

La fel ca universul nostru, viața a devenit treptat tot mai complexă și mai interesantă,* iar după cum voi arăta, găsesc util să clasificăm formele de viață în trei niveluri de complexitate: Viața 1.0, 2.0 și 3.0. Aceste trei niveluri sunt prezentate schematic în figura 1.1.

Nu știm deocamdată cum, când și unde a apărut pentru prima dată viața în universul nostru, dar există dovezi convingătoare că aici, pe Pământ, viața a apărut cu vreo 4 miliarde de ani în urmă. Nu după multă vreme pe planeta noastră forfotea o mare bogăție de forme de viață. Cele de succes, care curând și-au surclasat competitorii, au putut cumva reacționa la mediul lor. Mai precis, ele au fost ceea ce informaticienii numesc „agenți inteligenți”: entități care culeg informații despre mediu de la senzori,

* De ce a devenit viața tot mai complexă? Evoluția răsplătește viața care e suficient de complexă pentru a prezice și exploata regularitățile din mediul ei, așa încât într-un mediu mai complex va evolua o viață mai complexă și mai inteligentă. Această viață mai deșteaptă creează un mediu mai complex pentru formele de viață aflate în competiție, care, la rândul lor, evoluează devenind mai complexe, creând în cele din urmă ecosistemul unei vieți extrem de complexe. (N. a.)

iar apoi procesează informațiile pentru a hotărî cum să reacționeze la mediu. Asta poate include o procesare foarte complexă a informației, ca atunci când folosești informația primită de la ochi și urechi pentru a hotărî ce să spui într-o conversație. Dar poate include și un hardware și software destul de simple.

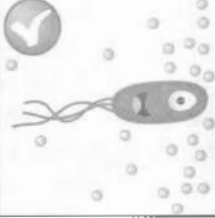
Poate să-și proiecteze hardware-ul?			 
Poate să-și proiecteze software-ul?		 	 
Poate supraviețui și se poate replica?	 	 	 
	Viața 1.0 (biologică simplă)	Viața 2.0 (culturală)	Viața 3.0 (tehnologică)

Figura 1.1 Cele trei etape ale vieții: evoluție biologică, evoluție culturală și evoluție tehnologică. Viața 1.0 e incapabilă să-și reproiecteze hardware-ul sau software-ul în timpul vieții: ambele sunt determinate de ADN-ul ei, și se modifică doar prin evoluția de-a lungul a numeroase generații. Dimpotrivă, Viața 2.0 poate reproiecta mult din software-ul ei: oamenii pot căpăta noi deprinderi complexe – de pildă, limbaje, sporturi sau profesii – și-și pot actualiza perspectiva asupra lumii și obiectivele. Viața 3.0, care nu există încă pe Pământ, își poate reproiecta spectaculos nu numai software-ul, dar și hardware-ul, în loc să aștepte ca acesta să evolueze treptat de-a lungul generațiilor.

De pildă, multe bacterii au un senzor care măsoară concentrația de glucide din lichidul din jurul lor, și pot înota folosind structuri propulsoare numite flageli. Hardware-ul care leagă senzorul de flageli poate pune în aplicare următorul algoritm simplu, dar util: „Dacă senzorul meu pentru concentrația glucidelor raportează o valoare mai mică decât cu câteva secunde în urmă, atunci inversează rotația flagelilor mei, așa încât să schimb direcția.”

Ați învățat să vorbiți și nenumărate alte lucruri. Bacteriile însă nu prea se pricep la învățat. ADN-ul lor le proiectează nu doar hardware-ul (senzorii pentru glucide, flagelii), ci și software-ul. Ele nu învață niciodată să înoate spre glucide; în schimb, acel algoritm a fost codificat în ADN-ul lor de la început. A existat, desigur, un fel de proces de învățare, dar el n-a avut loc în cursul vieții unei anumite bacterii, ci s-a desfășurat în evoluția anterioară a acelei specii de bacterii, printr-un îndelungat șir de încercări și eșecuri de-a lungul multor generații, în care selecția naturală a favorizat acele mutații întâmplătoare din ADN care îmbunătățeau consumul de glucide. Unele mutații au îmbunătățit proiectarea flagelilor și a altor componente ale hardware-ului, iar altele au îmbunătățit sistemul de procesare a informației care aplică algoritmul de găsimă a glucidelor și alte programe.

Asemenea bacterii sunt un exemplu pentru ceea ce numesc „Viața 1.0”: *viața în care hardware-ul și software-ul sunt rezultatul evoluției, nu sunt proiectate*. Prin software înțeleg toți algoritmi și toate cunoștințele pe care le folosești pentru a procesa informația primită de la simțuri și hotărăști ce să faci – totul, de la capacitatea de a-ți recunoaște prietenii când îi vezi, până la cea de a merge, citi, scrie, calcula, cânta și spune bancuri.

Nu erai în stare să îndeplinești nici una dintre aceste sarcini când te-ai născut, deci întreg acest software a fost programat în creierul tău mai târziu, prin procesul pe care îl numim învățare. Deși planul de învățare din copilărie e în mare parte determinat de familie și profesori, care hotărăsc ce trebuie să asimilezi, treptat capeți tot mai multă putere pentru a-ți proiecta software-ul. Poate că școala îți permite să alegi o limbă străină: Vrei să instalezi în creierul tău un modul software care te ajută să vorbești engleza,

sau unul care te ajută să vorbești franceza? Vrei să înveți să joci tenis sau șah? Vrei să studiezi pentru a deveni avocat sau farmacist? Vrei să afli mai multe despre inteligența artificială (IA) și viitorul vieții citind o carte pe tema asta?

Capacitatea Vieții 2.0 de a-și proiecta software-ul îi permite să fie mult mai deșteaptă decât Viața 1.0. O inteligență superioară presupune deopotrivă mult hardware (constând în atomi) și mult software (constând în biți). Faptul că majoritatea hardware-ului uman e adăugat după naștere (prin creștere) este avantajos, căci dimensiunea la care ajungi în cele din urmă nu e determinată de lărgimea colului uterin al mamei. De asemenea, faptul că majoritatea software-ului nostru uman e adăugat după naștere (prin învățare) este avantajos, căci inteligența la care ajungem în cele din urmă nu e limitată de cantitatea de informație care ne poate fi transmisă în momentul conceperii prin ADN, în stilul Vieții 1.0. Eu cântăresc de vreo douăzeci și cinci de ori mai mult decât atunci când m-am născut, iar conexiunile sinaptice care leagă neuronii din creierul meu pot stoca de vreo sută de mii de ori mai multă informație decât ADN-ul cu care m-am născut. Sinapsele stochează toate cunoștințele și deprinderile tale însumând aproximativ 100 de terabyți de informație, în vreme ce ADN-ul stochează doar aproximativ un gigabyte, care abia ajunge pentru un singur film. Este deci fizic imposibil ca un prunc să se nască vorbind perfect limba maternă și gata să dea examenul de admitere la liceu: informația nu poate fi pre-încărcată în creierul lui, fiindcă modulul de informație principal primit de la părinți (ADN-ul lui) nu are capacitatea suficientă de stocare a informației.

Înzestrarea de a-și proiecta software-ul îi permite Vieții 2.0 să fie nu doar mai deșteaptă decât Viața 1.0, dar și mai flexibilă. Dacă mediul se schimbă, 1.0 se poate adapta numai evoluând lent de-a lungul a numeroase generații. Viața 2.0 se poate însă adapta aproape instantaneu, prin actualizarea software-ului. De pildă, bacteriile care se confruntă frecvent cu antibiotice pot dezvolta rezistență la medicamente în cursul multor generații, dar o bacterie individuală nu-și schimbă comportamentul; în schimb, o fată care află că are alergii la alune își va modifica imediat comportamentul, evitând

alunele. Această flexibilitate oferă Vieții 2.0 un avantaj încă și mai mare la nivelul populației: dacă informația din ADN-ul uman n-a evoluat spectaculos în ultimii 50 000 de ani, informația stocată colectiv în creierele noastre, în cărți și calculatoare a explodat. Prin instalarea unui modul software care ne permite să comunicăm prin limbajul verbal sofisticat, am făcut astfel încât majoritatea informației utile stocată în creierul unei persoane să poată fi copiată în alte creiere, supraviețuind morții creierului original. Prin instalarea unui modul software care ne permite să citim și să scriem, am putut stoca și deține în comun mult mai multă informație decât sunt în stare să memoreze oamenii. Prin dezvoltarea unui software cerebral capabil să producă tehnologie (studiind științele și ingineria), am făcut ca mare parte din informație să poate fi accesată de aproape oricine cu doar câteva clicuri.

Această flexibilitate a permis Vieții 2.0 să domine Pământul. Eliberate de cătușele geneticii, cunoștințele însumate ale omenirii au sporit accelerat, fiecare realizare înlesnind-o pe următoarea: limbajul, scrierea, tiparul, știința modernă, calculatoarele, internetul etc. Această evoluție culturală tot mai rapidă a software-ului nostru comun a devenit forța dominantă care ne făurește viitorul, făcând ca lenta evoluție biologică să devină irelevantă.

În ciuda celor mai puternice tehnologii de care dispunem astăzi, toate formele de viață cunoscute rămân însă fundamental limitate de hardware-ul lor biologic. Nici una nu poate trăi un milion de ani, nu poate memora întreaga Wikipedie, înțelege întreaga știință la care am ajuns și nu poate zbura în spațiu fără o navă. Nici una nu poate transforma cosmosul nostru în mare parte lipsit de viață într-o biosferă diversă care va înflori miliarde sau bilioane de ani, permițând universului să-și împlinească în sfârșit posibilitățile și să se deștepte pe deplin. Pentru toate acestea e nevoie ca viața să facă încă un pas înainte, către Viața 3.0, care poate proiecta nu doar software-ul, dar și hardware-ul ei. Cu alte cuvinte, Viața 3.0 e stăpâna propriului ei destin, eliberată în cele din urmă complet de cătușele evoluției.

Granițele dintre cele trei etape ale vieții nu sunt precis trasate. Dacă bacteriile reprezintă Viața 1.0, iar oamenii Viața 2.0, atunci

pe șoareci i-am putea clasifica drept 1.1: ei pot învăța o sumedenie de lucruri, dar nu suficiente pentru a dezvolta limbajul sau a inventa internetul. În plus, fiindcă le lipsește limbajul, ceea ce învață se pierde atunci când mor, nu e transmis generației următoare. În mod asemănător, am putea spune că oamenii de azi ar trebui socotiți Viața 2.1: putem realiza mici upgradări ale hardware-ului, cum sunt implantarea dinților și genunchilor artificiali, stimulatorilor cardiace, dar nu și ceva atât de spectaculos cum ar fi să devenim de zece ori mai înalți sau să dobândim un creier de o mie de ori mai mare.

Pe scurt, putem împărți dezvoltarea vieții în trei etape, diferențiate prin capacitatea vieții de a se proiecta pe sine însăși:

- Viața 1.0 (etapa biologică): își evoluează hardware-ul și software-ul
- Viața 2.0 (etapa culturală): își evoluează hardware-ul, își proiectează mult din software
- Viața 3.0 (etapa tehnologică): își proiectează hardware-ul și software-ul

După 13,8 miliarde de ani de evoluție cosmică, dezvoltarea s-a accelerat spectaculos aici, pe Pământ: Viața 1.0 a apărut acum vreo patru miliarde de ani, Viața 2.0 (noi, oamenii) a apărut cu vreo sută de mii de ani în urmă, iar mulți cercetători IA cred că Viața 3.0 ar putea apărea în cursul următorului secol, eventual chiar în timpul vieții noastre, adusă de progresele IA. Ce anume se va întâmpla și ce va însemna asta pentru noi? Acesta e subiectul cărții de față.

CONTROVERSE

Întrebările de mai sus iscă numeroase controverse, specialiștii în IA contrazicându-se aprig nu numai în predicțiile lor, dar și în reacțiile lor afective, care variază de la optimism încrezător la îngrijorare acută. Nu există consens nici măcar în privința efectelor economice, juridice și militare pe termen scurt, iar dezacordul crește atunci când extindem orizontul de timp și ne punem

problema *inteligenței artificiale generale* (IAG) – în special în legătură cu capacitatea IAG de a atinge nivelul uman și de a-l depăși, permițând apariția Vieții 3.0. *Inteligența generală* poate îndeplini practic orice obiectiv, inclusiv învățarea, spre deosebire de inteligența limitată a unui program pentru jocul de șah.

Controversa privind Viața 3.0 se concentrează în jurul a două întrebări: când și ce? Când se va întâmpla (dacă se va întâmpla) și ce va însemna asta pentru omenire? Cred că există trei școli de gândire care trebuie luate în serios, pentru că fiecare include experți de nivel mondial. După cum se vede în figura 1.2, îi numesc *utopiști digitali*, *tehnosceptici* și *membri ai mișcării IA benefică*. Dați-mi voie să vă prezint pe unii dintre cei mai elocvenți exponenți ai lor.

UTOPIȘTI DIGITALI

Pe când eram puști mi-i închipuiam pe miliardari plini de morgă și aroganță. Când l-am întâlnit pentru prima oară la Google în 2008, Larry Page mi-a spulberat complet aceste clișee. Îmbrăcat în jeanși și cu un tricou modest, și-ar fi găsit locul la un picnic de la MIT (Massachusetts Institute of Technology – Institutul Tehnologic din Massachusetts). Vocea lui domoală și zâmbetul prietenos m-au făcut să mă simt relaxat. Pe 18 iulie 2015 ne-am întâlnit în Napa Valley la o petrecere dată de Elon Musk și soția sa de pe atunci, Talulah, și am început o discuție despre pornirile scatalogice ale copiilor noștri. I-am recomandat cartea clasică în domeniu, *The Day My Butt Went Psycho* de Andy Griffith, iar Larry a comandat-o pe loc. Aproape uitasem că el ar putea intra în istorie drept cel mai influent om din toate timpurile: dacă viața digitală superinteligentă va apărea în timpul vieții mele, cred că asta se va datora deciziilor lui Larry.

Am cinat apoi împreună cu soțiile noastre, Lucy și Meia, vorbind despre caracterul inevitabil al dobândirii conștiinței de către mașini – pentru el, o falsă problemă. Mai târziu, după cocteiluri, el și Elon au discutat însuflețit despre viitorul IA și ce trebuie făcut.

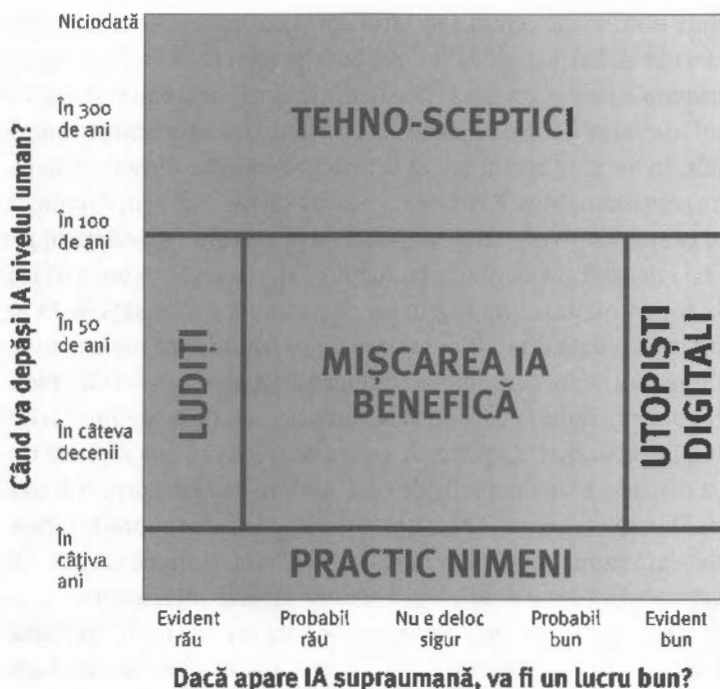


Figura 1.2 Majoritatea controverselor legate de inteligența artificială (dacă poate egala omul în orice sarcină cognitivă) se concentrează în jurul a două probleme: când se va întâmpla (dacă se va întâmpla) și dacă va fi un lucru bun pentru omenire. Tehno-scepticii și utopiștii digitali sunt de acord că n-ar trebui să ne batem capul cu ele, dar din motive diferite: primii sunt convinși că inteligența artificială generală (IAG) la nivel uman nu va apărea în viitorul previzibil, în vreme ce ultimii cred că va apărea și va fi fără îndoială un lucru bun. Mișcarea IA benefică susține că îngrijorarea e îndreptățită și necesară, fiindcă cercetările și dezbaterile privind siguranța IA sporesc acum șansele unui rezultat bun. Ludiții* sunt convinși că rezultatul va fi rău și se opun inteligenței artificiale. Figura e parțial inspirată de Tim Urban.¹

* Ludiții (*luddites*) au fost membrii unei organizații secrete a textiliștilor din Anglia secolului XIX, care distrugeau mașinile de țesut, temându-se că munca lor manuală va fi înlocuită de acestea. În sens general, e vorba despre cei care, simțindu-se amenințați, se opun progresului tehnologic. (N. t.)

Spre dimineață, cercul spectatorilor și chibiților continua să se lărgască. Larry a pledat cu pasiune pentru ceea ce eu numesc *utopism digital*: viața digitală este următorul pas natural și dezirabil în evoluția cosmică, iar dacă noi vom lăsa libere mințile digitale, în loc să le oprim sau să le înrobim, rezultatul va fi aproape cu certitudine bun. Îl consider pe Larry cel mai influent exponent al utopismului digital. Dacă viața se va răspândi vreodată în galaxia noastră și dincolo de ea, susținea el, atunci ar trebui s-o facă în formă digitală. Era îngrijorat că paranoia declanșată de IA ar întârzia utopia digitală și/sau ar face ca militarii să preia controlul asupra IA, în contradicție cu sloganul Google „Nu fiți răi”. Elon replica rugându-l pe Larry să lămurească detaliile argumentelor lui, întrebându-l, de pildă, de ce era convins că viața digitală nu va distruge toate lucrurile de care suntem atașați. Larry îl acuza pe Elon că e „specist” – că tratează drept inferioare anumite forme de viață numai fiindcă se bazează pe siliciu, și nu pe carbon. În capitolul 4 vom reveni asupra acestor aspecte interesante.

Deși în acea noapte călduroasă de vară cei mai mulți îl combăteau pe Larry, utopismul digital apărat atât de elocvent de el are numeroși susținători de frunte. Roboticianul și futurologul Hans Moravec a inspirat o întreagă generație de utopiști digitali cu clasică sa carte din 1988 *Mind Children*, tradiție continuată și amplificată de inventatorul Ray Kurzweil. Richard Sutton, unul dintre pionierii domeniului IA numit „învățare prin consolidare” (*reinforcement learning*), a pledat cu pasiune în favoarea utopismului digital la o conferință din Puerto Rico despre care voi vorbi în curând.

Tehno-sceptici

Gânditorii aparținând unui alt grup important nu-și fac nici ei griji în privința IA, dar dintr-un motiv diferit: crearea IAG supraumane li se pare atât de greu de îndeplinit în următoarele sute de ani, încât e ridicol să te preocupe acum. Această perspectivă tehno-sceptică e limpede exprimată de Andrew Ng: „Să te temi de roboți uci-gași e ca și cum te-ai teme de suprapopularea planetei Marte.”

Andrew a fost director de cercetare la Baidu, acel Google chinezesc, și și-a reluat argumentele când ne-am întâlnit la o conferință la Boston. Mi-a spus și că îngrijorarea legată de riscul IA ar putea fi o diversivune primejdioasă, încetinind progresele IA. Idei asemănătoare sunt împărtășite și de Rodney Brooks, fost profesor la MIT, creatorul aspiratorului robotic Roomba și al robotului industrial Baxter. E interesant de observat că deși utopiștii digitali și tehnoscepticii cad de acord asupra faptului că IA nu trebuie să ne îngrijoreze, altminteri au puține în comun. Majoritatea utopiștilor cred că IAG la nivel uman ar putea apărea într-un interval de timp între douăzeci și o sută de ani, ceea ce pentru tehnosceptici sunt vise deșarte, exaltări ale computeriștilor. Când l-am întâlnit pe Rodney Brooks la o petrecere în decembrie 2014, mi-a spus că era 100% convins că nu se va întâmpla în timpul vieții mele. „Sunteți sigur că nu vreți să spuneți 99%?” l-am întrebat într-un e-mail, la care mi-a răspuns: „Nu 99%, 100%. Pur și simplu n-o să se întâmple.”

Mișcarea IA benefică

Când l-am întâlnit pentru prima oară pe Stuart Russell, în iunie 2014 într-o cafenea pariziană, mi-a apărut ca exponentul tipic al unui gentleman britanic. Elocvent, grav, vorbind calm, dar cu o lucire cutezătoare în ochi, era întruchiparea modernă a lui Phileas Fogg, eroul copilăriei mele din romanul lui Jules Verne *Ocolul Pământului în optzeci de zile*. Deși era unul dintre cei mai celebri cercetători IA, coautor al cărții de referință în domeniu, modestia și căldura sa m-au făcut să mă simt în largul meu. Mi-a spus că progresele IA îl convinseseră că IAG la nivel uman poate fi obținută în acest secol, dar că un rezultat benefic nu putea fi garantat. Trebuia mai întâi să răspundem la întrebări esențiale, iar ele sunt atât de dificile, încât ar trebui să le cercetăm încă de-acum pentru a avea răspunsurile când va fi nevoie de ele.

Ideile lui Stuart sunt astăzi predominante, iar multe grupuri din lumea întreagă se ocupă cu studiul siguranței IA pentru care

pledează el. Însă lucrurile n-au stat întotdeauna așa. Într-un articol din *Washington Post* se spunea că 2015 a fost anul în care siguranța IA a intrat în centrul atenției. Înainte, dacă vorbeai despre riscurile IA, erai tratat drept ludic care se opune progresului IA. După cum vom vedea în capitolul 5, temeri asemănătoare celor ale lui Stuart au fost exprimate pentru prima dată cu mai bine de o jumătate de secol în urmă de Alan Turing, părintele calculatului, și matematicianul Irving J. Good, care a lucrat cu Turing la spargerea codurilor germane în timpul celui de-al Doilea Război Mondial. În deceniul precedent, asemenea preocupări aveau doar câțiva gânditori independenți care nu erau specialiști în IA – de pildă, Eliezer Yudkowsky, Michael Vassar și Nick Bostrom –, fără ecou în rândul specialiștilor, care se concentrau asupra sarcinilor curente de îmbunătățire a sistemelor IA. Dintre cercetătorii IA care nutreau anumite îngrijorări, mulți ezitau să le dea glas de teamă că vor fi catalogați drept tehnofobi alarmiști.

Simțeam că această polarizare trebuia depășită, așa încât întreaga comunitate IA să participe la dezbateră privind felul în care poate fi construită IA benefică. Din fericire, nu eram singurul. În primăvara lui 2014 am înființat o organizație non-profit numită Institutul pentru Viitorul Vieții (Future of Life Institute, FLI; <http://futureoflife.org>) împreună cu soția mea Meia, prietenul meu fizician Anthony Agguire, Viktoriya Krakovna, absolventă de Harvard, și Jaan Tallinn, fondatorul Skype-ului. Scopul nostru era simplu: să contribuim la asigurarea viitorului vieții, un viitor cât mai bun cu putință. Mai exact, simțeam că tehnologia dădea vieții puterea fie de a cunoaște o înflorire fără precedent, fie de a se autodistruge, iar noi preferam prima variantă.

A avut loc mai întâi o ședință de *brainstorming* la noi acasă, pe 15 martie 2014, la care au participat vreo treizeci de studenți, profesori și persoane interesate de subiect din zona Bostonului. Am căzut de acord asupra faptului că, deși trebuie să ne preocupe biotehnologiile, armele nucleare și modificarea climei, principala noastră obiectiv este aducerea în prim-plan a cercetărilor privind siguranța IA. Colegul meu de la MIT Franck Wilczek, care a

primit Premiul Nobel pentru contribuții la fizica cuarcilor, a propus să începem prin a scrie un editorial de opinie în care să atragem atenția asupra temei, făcând-o mai greu de ignorat. Am luat legătura cu Stuart Russell (pe care nu-l întâlnisem încă) și Stephen Hawking, iar amândoi au fost de acord să fie coautori, alături de mine și de Franck. Editorialul ne-a fost respins de *The New York Times* și de multe alte ziare americane, așa că l-am postat pe contul meu de blog de la *Huffington Post*. Spre bucuria mea, Arianna Huffington însăși mi-a trimis un mesaj în care spunea „Sunt încântată să-l avem! Îl vom posta la #1”, iar această plasare la începutul paginii a declanșat un val de acoperire media a siguranței IA care a durat tot anul, cu intervențiile lui Elon Musk, Bill Gates și ale altor figuri proeminente din tehnologie. Cartea lui Nick Bostrom *Superintelligence* a apărut în acea toamnă, întetind dezbaterea publică.

Următorul obiectiv al campaniei noastre pentru IA benefică a fost să-i reunim pe principalii cercetători IA într-o conferință la care să se lămurească neînțelegerile, să se ajungă la un consens și să se avanseze planuri constructive. Știam că e greu de convins un asemenea grup de persoane ilustre să vină la o conferință organizată de outsideri pe care nu-i cunoșteau, mai ales că tema era controversată, așa că ne-am străduit pe cât am putut – am interzis accesul presei, am ales ca loc de desfășurare o stațiune din Puerto Rico, n-am cerut taxe de participare (grație generozității lui Jaan Tallinn) și am găsit cel mai nealarmist titlu imaginabil: „Viitorul IA: Posibilități și provocări”. În plus, ne-am asociat cu Stuart Russell, mulțumită căruia am lărgit comitetul de organizare pentru a include specialiști IA din mediul academic și din industrie – între care, Demis Hassabis de la DeepMind, Google, care a arătat că IA îi poate învinge pe oameni chiar și în jocuri precum go. Cunoșcându-l pe Demis, mi-am dat seama că avea ambiția de a face IA nu doar puternică, dar și benefică.

Urmarea a fost o remarcabilă întrunire a unor mari minți (figura 1.3). Cercetătorii IA s-au întâlnit cu economiști și juriști de vârf, lideri din tehnologie (între care Elon Musk) și alți gânditori



Figura 1.3 Conferința din ianuarie 2013 din Puerto Rico a reunit un grup remarcabil de cercetători IA și din domeniile conexe. Pe rândul din spate, de la stânga la dreapta: Tom Mitchell, Seán Ó hÉigeartaigh, Huw Price, Shamil Chandaria, Jaan Tallinn, Stuart Russell, Bill Hibbard, Blaise Agüera y Arcas, Anders Sandberg, Daniel Dewey, Stuart Armstrong, Luke Muehlhauser, Tom Dietterich, Michael Osborne, James Manyika, Ajay Agrawal, Richard Mallah, Nancy Chang, Matthew Putman. Alții, stând în picioare, de la stânga la dreapta: Marilyn Thompson, Rich Sutton, Alex Wissner-Gross, Sam Teller, Toby Ord, Joscha Bach, Katja Grace, Adrian Weller, Heather Roff-Perkins, Dileep George, Shane Legg, Demis Hassabis, Wendell Wallach, Charina Choi, Ilya Sutskever, Kent Walker, Cecilia Tilli, Nick Bostrom, Erik Brynjolfsson, Steve Crossan, Mustafa Suleyman, Scott Phoenix, Neil Jacobstein, Murray Shanahan, Robin Hanson, Francesca Rossi, Nate Soares, Elon Musk, Andrew McAfee, Bart Selman, Michele Reilly, Aaron VanDevender, Max Tegmark, Margaret Boden, Joshua Greene, Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel, James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva-Melissa Tez, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. În rândul din față: Sam Harris, Tomaso Poggio, Marin Soljačić, Viktoriya Krakovna, Meia Chita-Tegmark. În spatele aparatului de fotografiat: Anthony Aguirre (introdus în imagine prin photoshop).

(între care Vernor Vinge, cel care a introdus termenul „singularitate”, tratat în capitolul 4). Rezultatul a depășit cele mai optimiste așteptări. Poate că a fost o combinație între vin și soare, sau poate că pur și simplu sosise momentul potrivit: în ciuda subiectului controversat, s-a ajuns la un consens remarcabil, care a fost



Figura 1.4 Deși s-a spus adesea că Elon Musk se află în dispută cu comunitatea IA, s-a ajuns la un consens larg privind necesitatea cercetărilor de siguranță IA. În fotografia făcută pe 4 ianuarie 2015, Tom Dietterich, președintele Asociației pentru Progresul Inteligenței Artificiale împărtășește cu Elon entuziasmul pentru noul program de cercetare în siguranța IA pe care Elon tocmai promisese să-l finanțeze. În spatele lor se zăresc Meia-Cita Tegmark și Viktoriya Kravovna, fondatoare ale FLI.

exprimat sub forma unei scrisori deschise² semnate în cele din urmă de peste opt mii de oameni, incluzând un veritabil „who’s who” în IA. În esență, scrisoarea spunea că IA trebuia îndrumată pentru a deveni benefică, și menționa în detaliu temele de cercetare care urmau să fie abordate în vederea acestui obiectiv. Mișcarea pentru IA benefică intrase în prim-plan. Evoluția ei ulterioară va fi prezentată în cartea de față.

Altă concluzie a conferinței a fost că problemele ridicate de succesul IA nu sunt doar fascinante în plan intelectual, ci au și o importanță crucială în plan moral, fiindcă alegerile noastre pot afecta viitorul vieții. Alegerile omenirii din trecut au avut uneori o mare însemnătate morală, dar ea a fost întotdeauna limitată: ne-am revenit după dezastre, iar până și imperiile cele mai falnice s-au prăbușit în cele din urmă. Generațiile din trecut știau că, așa cum soarele va răsări a doua zi, și oamenii vor avea a doua zi de înfruntat nenorociri precum sărăcia, bolile și războiul. Unii conferențieri din Puerto Rico au susținut însă că de data asta lucrurile

ar putea sta altfel: pentru prima oară am putea crea o tehnologie suficient de puternică pentru a pune capăt definitiv acestor nenorociri – sau pentru a pune capăt omenirii înseși. Am putea făuri societăți care să prospere ca niciodată, sau un stat global kaskian care supraveghează totul, atât de puternic încât să fie imposibil de răsturnat.

CONFUZII

Când am plecat din Puerto Rico eram convins că discuțiile despre viitorul IA trebuie să continue, fiindcă e cea mai importantă dezbateră a epocii noastre.* Este dezbateră despre viitorul colectiv al nostru, al tuturor, deci n-ar trebui să se limiteze la cercetătorii IA. Am scris cartea de față în speranța că tu, dragă cititorule, vei participa la această dezbateră. Ce fel de viitor îți dorești? Vrei să construim arme autonome letale? Ce-ai vrea să se întâmple cu automatizarea muncii? Ce carieră i-ai sfătui să urmeze pe copiii de azi? Preferi ca noi joburi să le înlocuiască pe cele vechi, sau o societate fără joburi, în care toată lumea să se relaxeze, profitând de bogăția produsă de mașini? Privind mai departe, ai vrea să creăm Viața 3.0 și s-o răspândim în univers? Mașinile inteligente ne vor înlocui, vor coexista cu noi sau se vor contopi cu noi? Ce va însemna să fii om, în epoca inteligenței artificiale? Ce ai vrea tu să însemne și cum putem face ca viitorul să arate așa cum îți dorești?

Scopul cărții de față este de a te invita să participi la dezbateră. După cum am spus, sunt dezbateri fascinante în care mari experți se află pe poziții opuse, dar am văzut și numeroase exemple de

* Dezbateră despre IA e importantă atât pentru urgența, cât și pentru impactul ei. În comparație cu modificarea climei, care va avea efecte dăunătoare peste cincizeci sau peste două sute de ani, mulți experți cred că IA va avea un impact mai puternic în următoarele decenii – și ne-ar putea oferi tehnologia pentru a stăvili modificarea climei. În comparație cu problemele legate de război, terorism, șomaj, sărăcie, migrație și nedreptatea socială, apariția IA va avea un impact global mai mare – iar în cartea de față vom vedea ce efecte, bune sau rele, ar putea avea asupra tuturor acestor probleme. (N. a.)

Rezumat al terminologiei:

Viață	Proces care își păstrează complexitatea și se replicează
Viața 1.0	Viața care își evoluează hardware-ul și software-ul (etapa biologică):
Viața 2.0	Viața care își evoluează hardware-ul, dar își proiectează mult din software (etapa culturală)
Viața 3.0	Viața care își proiectează hardware-ul și software-ul (etapa tehnologică)
Inteligență	Capacitatea de a îndeplini obiective complexe
Inteligență artificială (IA)	Inteligență nebiologică
Inteligență limitată	Capacitatea de a îndeplini un set limitat de sarcini (de ex., de a juca șah, de a conduce o mașină)
Inteligență generală	Capacitatea de a îndeplini practic orice obiectiv, inclusiv învățarea
Inteligență universală	Capacitatea de a dobândi inteligență generală, având acces la date și resurse
Inteligență artificială generală (la nivel uman) (IAG)	Capacitatea de a îndeplini orice sarcină cognitivă, cel puțin la fel de bine ca oamenii
IA la nivel uman	IAG
IA tare	IAG
Superinteligență	Inteligență generală mult dincolo de nivelul uman
Civilizație	Grup în care interacționează forme inteligente de viață
Conștiință	Experiență subiectivă
Qualia	Expresii individuale ale experienței subiective
Etică	Principii care ne spun cum trebuie să ne comportăm
Teleologie	Explicarea lucrurilor pe baza obiectivelor sau scopurilor lor, nu pe baza cauzelor
Comportament orientat spre obiective	Comportament mai ușor de explicat prin efectele lui decât prin cauzele lui
A avea un obiectiv	A manifesta un comportament orientat spre obiective
A avea un scop	A sluji obiectivele proprii sau ale altei entități
IA prietenoasă	Superinteligență ale cărei obiecte sunt alinate cu ale noastre
Cyborg	Hibrid om-mașină
Explozie a inteligenței	Autoperfecționare recursivă rapidă conducând spre super-inteligență
Singularitate	Explozie a inteligenței
Univers	Regiunea spațiului din care lumina a avut timp să ajungă la noi în cursul celor 13,8 miliarde de ani scurși de la big bang

Tabelul 1.1 Numeroase confuzii legate de IA sunt provocate de sensurile diferite pe care le dăm termenilor. Iată ce sens le dau eu în cartea de față. (Unele definiții vor fi explicate în capitolele următoare.)

pseudodezbateri confuze plictisitoare. Pentru a ne concentra asupra discuțiilor interesante și întrebărilor deschise, iar nu asupra confuziilor, e cazul să lămurim cele mai frecvente neînțelegeri.

Există multe definiții pentru termeni precum „viață”, „intelligență” și „conștiință”, iar multe neînțelegeri sunt provocate de faptul că oamenii folosesc un cuvânt în sensuri diferite. Pentru a nu cădea în această capcană, am notat în tabelul 1.1 sensul în care folosesc eu termenii-cheie. Unele definiții vor fi explicitate în capitolele următoare. Nu pretind că definițiile mele sunt mai bune decât ale altora, vreau doar să risipesc confuziile. Observați că în general folosesc definiții cuprinzătoare, care evită antropocentrismul și pot fi aplicate mașinilor la fel de bine ca oamenilor. Vă invit să citiți acum tabelul și să reveniți la el dacă sunteți derutați de sensul în care folosesc termenii, în special în capitolele 4–8.

În plus față de confuzia provocată de terminologie, multe debateri despre IA sunt afectate de neînțelegeri simple. Vă propun să le lămurim pe cele mai frecvente.

Miturile cronologiei

Primele neînțelegeri se referă la cronologia din figura 1.2: cât va dura până când mașinile vor depăși cu mult IAG la nivel uman? Ideea că am cunoaște răspunsul cu certitudine este greșită.

Un mit larg răspândit spune că vom ajunge la IAG supraumană în acest secol. De fapt, istoria e plină de previziuni exagerate legate de tehnologie. Unde sunt acele centrale de fuziune sau mașini zburătoare care ni s-au promis? Și în privința IA s-a exagerat în trecut, au făcut-o chiar unii dintre fondatorii domeniului. De pildă, John McCarthy (cel care a creat termenul „intelligență artificială”), Marvin Minsky, Nathaniel Rochester și Claude Shannon au prevăzut că aceasta va fi obținută în două luni cu calculatoare primitive: „Propunem ca un studiu asupra inteligenței artificiale să fie efectuat timp de două luni de către zece oameni în vara lui 1956 la Colegiul Dartmouth. [...] Se va încerca să se afle cum pot fi făcute mașinile să folosească limbajul, să formeze abstracțiuni și concepte, să rezolve probleme rezervate acum oamenilor și să

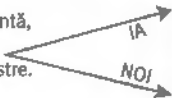
Mit În 2100 superinteligenta va fi inevitabilă. Mit În 2100 superinteligenta va fi imposibilă.	<table> <tr> <th>Mon</th><th>Tue</th><th>Wed</th><th>Thur</th><th>Fri</th><th>Sat</th><th>Sun</th></tr> <tr> <td></td><td></td><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr> <tr> <td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td></tr> <tr> <td>12</td><td>13</td><td>14</td><td>15</td><td>16</td><td>17</td><td>18</td></tr> <tr> <td>19</td><td>20</td><td>21</td><td>22</td><td>23</td><td>24</td><td>25</td></tr> <tr> <td>26</td><td>27</td><td>28</td><td>29</td><td>30</td><td></td><td></td></tr> </table>	Mon	Tue	Wed	Thur	Fri	Sat	Sun				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30			Fapt Poate apărea peste decenii, secole sau niciodată – experții IA se contrazic, iar noi pur și simplu nu putem ști. 
Mon	Tue	Wed	Thur	Fri	Sat	Sun																																						
			1	2	3	4																																						
5	6	7	8	9	10	11																																						
12	13	14	15	16	17	18																																						
19	20	21	22	23	24	25																																						
26	27	28	29	30																																								
Mit Numai ludiții sunt îngrijorați de IA. 	Fapt Mulți cercetători IA de vârf sunt preocupați. 																																											
Îngrijorare mitică IA devine rea. Îngrijorare mitică IA devine conștientă. 	Îngrijorare reală IA devine competentă, cu obiective diferite de ale noastre. 																																											
Mit Roboții sunt principala problemă. 	Fapt IA orientată diferit e principala problemă: nu are nevoie de corp, ci doar de conexiune internet. 																																											
Mit IA nu poate controla oamenii. 	Fapt Inteligenta permite controlul: noi controlăm tigrii pentru că suntem mai deștepți. 																																											
Mit Mașinile nu pot avea obiective. 	Fapt Rachetele termoghidate au un obiectiv. 																																											
Îngrijorare mitică Superinteligenta va sosi curând. 	Îngrijorare reală Va sosi cel puțin peste decenii, dar ar putea dura mult s-o facem sigură. 																																											

Figura 1.5 Mituri frecvent întâlnite privind IA superinteligentă.

se autoperfecționeze. Credem că într-una sau mai multe dintre aceste probleme se poate obține un progres important dacă un grup de oameni de știință atent selectați lucrează timp de o vară.”

Pe de altă parte, conform unui contra-mit răspândit, știm că *nu vom* ajunge la IAG supraumană în acest secol. Cercetătorii au făcut numeroase estimări privind timpul care ne separă de IAG supraumană, dar nu putem spune că probabilitatea pentru acest secol e zero, dat fiind jalnicul palmares al unor asemenea predicții tehnosceptice. De pildă, Ernest Rutherford, pesemne cel mai mare fizician nuclearist al epocii sale, a spus în 1933 – chiar înainte ca Leo Szilard să găsească lanțul de reacții nucleare – că energia nucleară e o prostie, iar în 1956 astronomul englez Richard Wooley a spus că discuțiile despre călătoria spațială sunt „vorbe în vânt”. Acest mit ia forma extremă în afirmația că nu se va ajunge niciodată la IAG supraumană, fiindcă e fizic imposibilă. Fizicienii știu totuși că un creier e alcătuit din cuarci și electroni aranjați pentru a acționa ca un calculator puternic, și că nici o lege a fizicii nu-i împiedică să se constituie într-o structură și mai inteligentă.

S-au făcut câteva sondaje printre cercetătorii IA, cu întrebarea când vom avea IAG la nivel uman cu o probabilitate de cel puțin 50%, iar în nici un sondaj experții n-au căzut de acord, așa că pur și simplu nu știm. De pildă, într-un asemenea sondaj făcut la conferința din Puerto Rico, răspunsul mediu a fost anul 2055, dar după unii cercetători va dura sute de ani sau mai mult.

Există și un mit care spune că cei preocupați de IA cred că aceasta va sosi foarte curând, când de fapt majoritatea estimează că vor mai trece cel puțin câteva decenii. Ei susțin însă că, nefiind 100% siguri că nu va fi în acest secol, e înțelept să începem încă de pe-acum cercetările legate de siguranță. După cum vom vedea în cartea de față, multe probleme de siguranță sunt atât de dificile, încât ar putea trece decenii până să fie rezolvate, și ar trebui s-o facem înainte ca niște programatori care beau Red Bull să se ho-tărăscă să treacă la IAG la nivel uman.

Miturile controversiei

Altă idee falsă este că numai niște ludiți ignoranți sunt preocupați de siguranța IA și pledează pentru cercetările în domeniu. Când Stuart Russell a menționat asta în cuvântarea lui din Puerto Rico, publicul a izbucnit în râs. Se mai spune și că sprijinul acordat cercetărilor în siguranța IA e controversată. De fapt, pentru a face o investiție modestă în cercetările privind siguranța IA nu-i nevoie să fii convins că riscurile sunt mari, la fel cum o investiție modestă în asigurarea locuinței e justificată pur și simplu de probabilitatea nenulă de a-ți lua foc casa.

Cred că mediile de informare au făcut ca dezbaterea despre siguranța IA să pară mai controversată decât e în realitate. În fond, teama se vinde bine, iar articolele care folosesc citate scoase din context pentru a proclama sfârșitul iminent primesc mai multe clicuri decât cele echilibrate și nuanțate. Prin urmare, doi oameni care au aflat unul despre poziția celuilalt doar din citate media vor crede probabil că se contrazic mai mult decât se întâmplă în realitate. De pildă un tehnosceptic care află despre poziția lui Bill Gates dintr-un tabloid britanic și-ar putea închipui că pentru acesta superinteligenta e iminentă. Apoi, un membru al mișcării IA benefică, necunoscând despre poziția lui Andrew Ng altceva decât afirmația menționată mai sus privind suprapopularea planetei Marte, ar putea crede că acestuia nu-i pasă de siguranța IA. Eu știu că-i pasă, dar cum Andrew crede că la IA se va ajunge într-un viitor îndepărtat, e interesat în primul rând de provocările IA pe termen scurt.

Mituri despre risc

Am citit exasperat acest titlu scris cu litere de-o șchioapă în *Daily Mail*³ „Stephen Hawking ne previne că ascensiunea roboților ar putea fi catastrofală pentru omenire.” Mi-a fost dat să văd nenumărate asemenea articole. De regulă, sunt însoțite de imaginea unui robot malefic purtând o armă, și sugerează că ar trebui să fim îngrijorați de roboți care se răscoală și ne ucid pentru că au devenit

conștienți și/sau răi. Asemenea articole rezumă tocmai scenariul care pe colegii mei din IA nu îi preocupă. El combină trei elemente: îngrijorarea față de conștiență, rău și roboți.

Când conduci mașina ai o experiență subiectivă a culorilor, sunetelor etc. Dar o mașină care se conduce singură are o asemenea experiență subiectivă? Simte ceva, sau e doar un zombi înconștient, fără experiențe subiective? Deși acest mister al conștiinței e interesant în sine, și vom vorbi despre el în capitolul 8, este irelevant pentru riscul IA. Dacă te calcă o mașină fără șofer, n-are importanță pentru tine dacă ea se simte în mod subiectiv conștientă. La fel, ce ne afectează pe noi, oamenii, este ce *face*, nu cum se simte în mod subiectiv IA.

Teama de mașini care devin rele e o altă falsă problemă. Adevăratul motiv de îngrijorare nu e răutatea, ci competența. O IA superinteligentă se pricepe prin definiție foarte bine să-și atingă obiectivele, oricare ar fi ele, așa că trebuie să ne asigurăm că obiectivele ei sunt aliniate cu ale noastre. Pesemne că nu ești un om care urăște furnicile și le calcă din răutate, dar dacă proiectezi o centrală hidroelectrică, iar un mușuroi de furnici se află în regiunea care va fi inundată, ghinionul lor. Mișcarea IA benefică urmărește să împiedice plasarea omenirii în locul acelor furnici.

Neînțelegerea legată de conștiință se leagă de mitul că mașinile nu pot avea obiective. Evident, mașinile pot avea obiective în sensul restrâns de a manifesta un comportament orientat spre obiective: comportamentul unei rachete termoghidate se explică cel mai simplu prin obiectivul de a lovi ținta. Dacă te simți amenințat de o mașină ale cărei obiective diferă de ale tale, atunci tocmai aceste obiective în sens restrâns te neliniștesc, nu dacă mașina e conștientă și percepe vreun scop. Dacă ești fugărit de o rachetă termoghidată, probabil că nu vei exclama „Nu-mi pasă, pentru că mașinile nu pot avea obiective!“.

Sunt de partea lui Rodney Brooks și a altor pionieri ai roboticii care se simt pe nedrept demonizați în tabloide fiindcă unii jurnaliști au făcut o fixație pentru roboți și-și împodobesc articolele cu monștri de metal cu ochi roșii strălucitori. De fapt, principala

preocupare a mișcării IA benefică nu sunt roboții, ci inteligența însăși – mai precis, inteligența ale cărei obiective diferă de ale noastre. Ea n-are nevoie de un corp robotic pentru a ne face rău, e suficientă o conexiune internet. În capitolul 4 vom vedea cum aceasta poate păcăli piețele financiare, poate întrece în iscusință cercetătorii umani, poate manipula lideri politici și dezvolta arme pe care noi nici măcar nu le-am înțelege. Chiar dacă i-ar fi imposibil fizic să construiască roboți, o IA superinteligentă și extrem de bogată ar putea cu ușurință plăti sau manipula milioane de oameni pentru a-i îndeplini înconștienți poruncile, ca în romanul SF al lui William Gibson *Neuromancer*.

Ideea greșită despre roboți se leagă de mitul că mașinile nu pot controla oamenii. Inteligența face cu puțință controlul: oamenii controlează tigrii nu pentru că sunt mai puternici, ci pentru că sunt mai deștepți. Asta înseamnă că, dacă încetăm să fim cei mai deștepți de pe planetă, s-ar putea și să pierdem controlul.

Figura 1.5 rezumă aceste idei false frecvent întâlnite, așa că ne putem debarasa de ele, pentru a ne concentra asupra controverSELor legitime deloc puține, după cum vom vedea.

DRUMUL DE URMAT

În restul acestei cărți vom explora împreună viitorul vieții cu IA. Vă propun să navigăm organizat prin acest subiect complex, explorând mai întâi, conceptual și cronologic, întreaga poveste a vieții, iar apoi să ne ocupăm de obiectivele și semnificația acțiunilor de urmat pentru a ne făuri viitorul dorit.

În capitolul 2 vom vorbi despre bazele inteligenței și felul în care materia neînsuflețită poate fi rearanjată pentru a memora, calcula și învăța. Îndreptându-ne spre viitor, povestea noastră se ramifică în mai multe scenarii definite de răspunsurile la anumite întrebări-cheie. Figura 1.6 rezumă întrebările-cheie pe care le vom întâlni pe măsură ce înaintăm în timp spre o IA tot mai avansată.

Încă de pe acum suntem confrunțați cu alegerea de a începe sau nu o cursă a înarmărilor IA și cu întrebarea ce putem face pentru ca sistemele IA de mâine să fie robuste și lipsite de erori. Dacă impactul economic al IA continuă să crească, trebuie să hotărâm cum să ne modernizăm legile și ce cariere să-i sfătuim pe copiii noștri să urmeze pentru a evita joburi care vor fi curând automatizate. Vom examina asemenea chestiuni cu bătaie scurtă în capitolul 3.

Dacă IA avansează până la nivel uman, trebuie să ne întrebăm cum ne putem asigura că e benefică și dacă e cazul să făurim o societate a relaxării care prosperă fără joburi. De aici se naște întrebarea dacă o explozie a inteligenței sau o lentă creștere constantă pot împinge IAG mult dincolo de nivelul uman. În capitolul 4 vom examina o gamă largă de asemenea scenarii, iar în capitolul 5 vom cerceta spectrul posibilităților care se deschid, de la cele distopice la cele utopice. Cine comandă – oamenii, IA sau cyborgii? Sunt oamenii tratați bine sau rău? Suntem oare înlocuiți, iar în acest caz, îi percepem pe înlocuitori drept învingători sau demni urmași? Sunt curios să știu pe care dintre scenariile din capitolul 5 îl preferați! Am creat un site, <http://AgeOfAi.org>, unde vă puteți spune părerea și intra în dialog.

În fine, vom înainta cu miliarde de ani în viitor în capitolul 6, unde, paradoxal, putem trage concluzii mai clare decât în capitolele precedente, căci limitele ultime ale vieții în cosmosul nostru nu sunt impuse de inteligență, ci de legile fizicii.

După ce vom fi încheiat explorarea istoriei inteligenței, ne vom confrunta cu întrebările ce viitor este de dorit și cum putem ajunge la el. Pentru a lega datele seci de problema scopului și a semnificației, vom examina bazele fizice ale obiectivelor în capitolul 7, iar conștiința în capitolul 8. În fine, în epilog, vom vedea ce poate fi făcut chiar acum pentru a contribui la edificarea viitorului pe care ni-l dorim.

Dacă îți place să citești pe sărite, majoritatea capitolelor sunt relativ autonome odată ce ai deprins terminologia și definițiile din acest prim capitol și de la începutul următorului. Dacă ești

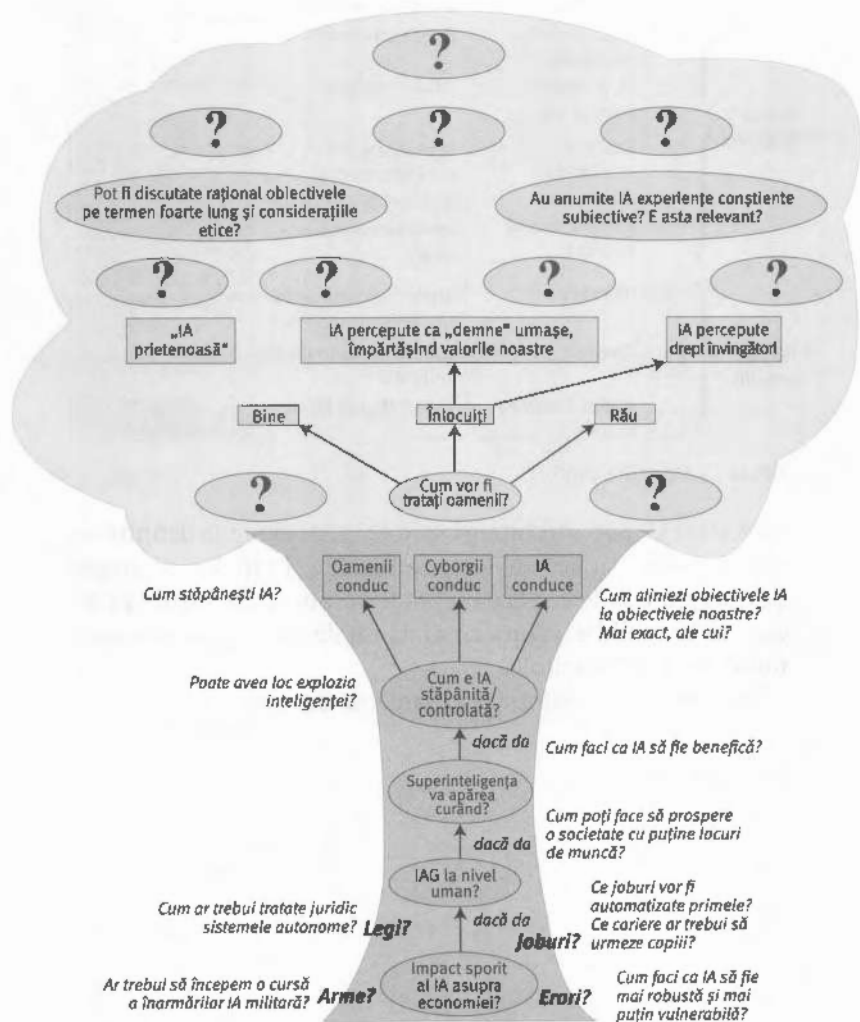


Figura 1.6 Întrebările interesante legate de IA depind de felul în care se ajunge la IA și de ramura pe care o va urma viitorul

	<i>Titlul capitolului prescurtat</i>	<i>Subiect</i>	<i>Caracter</i>
Istoria Inteligenței	Preludiu: Povestea Echipei Omega	Material pentru reflecție	Extrem speculativ
	1. Dezbateră	Idei-cheie, terminologie	Nu foarte speculativ
	2. Materia devine inteligentă	Bazele inteligenței	
	3. IA, economie, arme și legi	Viitorul apropiat	
	4. Explozia inteligenței?	Scenarii ale superinteligenței	Extrem speculativ
	5. Urmări	Următorii 10 000 de ani	
Istoria sensului	6. Zestrea noastră cosmică	Următoarele miliarde de ani	
	7. Obiective	Istoria comportamentu- lui orientat spre obiective	Nu foarte speculativ
	8. Conștiința	Conștiința naturală și artificială	Speculativ
	Epilog: Povestea Echipei FLI	Ce ar trebui să facem?	Nu foarte speculativ

Figura 1.7 Structura cărții

cercetător IA, poți sări întreg capitolul 2, cu excepția definițiilor de la început. Dacă ești novice, capitolele 2 și 3 îți demonstrează de ce capitolele de la 4 la 6 nu pot fi simplu etichetate drept po-vești SF imposibile. Figura 1.7 arată cât de speculativ este conți-nutul diverselor capitole.

Ne așteaptă o călătorie fascinantă. Să începem!

REZUMAT:

- Viața, definită ca proces care își păstrează și replichează complexitatea, poate cunoaște trei etape: etapa biologică (1.0), în care hardware-ul și software-ul ei evoluează, etapa culturală (2.0), în care își poate proiecta software-ul prin învățare, și etapa tehnologică (3.0), în care își poate proiecta și hardware-ul, devenind stăpână pe destinul ei.
- Inteligența artificială ne poate permite să lansăm Viața 3.0 în acest secol, și s-a declanșat o dezbatere fascinantă privind viitorul pe care ni-l dorim și calea de a ajunge la el. Trei tabere principale au intrat în dezbatere: tehnosceptici, utopiști digitali și mișcarea IA benefică.
- Tehnoscepticii cred că IAG supraumană e atât de dificil de creat, încât nu va apărea în următoarele sute de ani, deci e ridicol să ne preocupăm acum de ea (și de Viața 3.0).
- Utopiștii digitali cred că e probabilă apariția ei în acest secol și întâmpină cu inima deschisă Viața 3.0, considerând-o următorul pas, natural și de dorit, în evoluția cosmică.
- Mișcarea IA benefică crede de asemenea că e probabilă apariția ei în acest secol, dar nu consideră că rezultatul bun e garantat, ci că acesta trebuie asigurat printr-o muncă serioasă sub forma cercetării în siguranța IA.
- În afară de aceste controverse legitime, în care experții nu cad de acord, există și pseudocontroverse plictisitoare provocate de confuzii. De pildă, nu-ți pierde vremea discutând despre „viață”, „inteligență” sau „conștiință” până nu te asiguri că pentru tine și interlocutorul tău aceste cuvinte înseamnă același lucru! Cartea de față folosește definițiile din tabelul 1.1.
- Ferește-te și de afirmațiile false din figura 1.5: „În 2100 superinteligența va fi inevitabilă/imposibilă”, „Numai ludiții sunt îngrijorați de IA”, „Ne îngrijorează faptul că IA ar deveni rea și/sau conștientă, și că ne despart de ea doar câțiva ani”, „Roboții sunt principala problemă”, „IA nu poate controla oamenii și nu poate avea obiective”.
- În capitolele de la 2 la 6 vom explora povestea inteligenței de la umilele ei începuturi cu miliarde de ani în urmă până la posibilele viitoruri cosmice de peste miliarde de ani. Vom analiza întâi provocările pe termen scurt, legate de joburi, arme IA și încercarea de a ajunge la IAG la nivel uman, iar apoi un fascinant spectru de viitoruri posibile cu mașini inteligente și/sau oameni. Mă întreb ce variantă preferați!
- În capitolele de la 7 la 9 vom trece de la descrierea rece factuală la explorarea obiectivelor, conștiinței și sensului, și vom investiga ce putem face chiar acum pentru a contribui la edificarea viitorului dorit.
- Consider că această discuție despre viitorul vieții cu IA este cea mai importantă dezbatere a epocii noastre – așa încât vă invit să participați!

CAPITOLUL 2

Materia devine inteligentă

Hidrogenul [...], dacă are suficient timp la dispoziție, se transformă în oameni. — Edward Robert Harrison, 1995

Una dintre cele mai spectaculoase evoluții în cursul celor 13,8 miliarde de ani scurși de la big bang este că acea materie mută și neînsuflită a devenit inteligentă. Cum s-a petrecut asta și cum poate progresa inteligența în viitor? Ce are de spus știința despre istoria și destinul inteligenței în cosmosul nostru? Pentru a aborda aceste întrebări, vom examina în capitolul de față bazele inteligenței și elementele ei constitutive. Ce înseamnă că un bulgăre de materie e inteligent? Ce înseamnă că un obiect poate memora, calcula și învăța?

CE ESTE INTELIGENȚA?

Soția mea și cu mine am avut ocazia să asistăm la un simpozion despre inteligența artificială organizat de Fundația Suedeză Nobel, iar atunci când mai multor cercetători IA de vârf li s-a cerut să definească inteligența, s-au lansat într-o dispută lungă, fără să ajungă la consens. Ni s-a părut amuzant: nici măcar specialiștii în inteligența artificială nu cad de acord! Nu există așadar nici o definiție „corectă” indiscutabilă a inteligenței. Există însă multe definiții concurente, vizând, între altele, aptitudinea pentru

logică, înțelegere, planificare, cunoaștere afectivă, conștiință de sine, creativitate, rezolvare a problemelor și învățare.

În explorarea viitorului inteligenței, vrem să avem o perspectivă cât mai largă, fără să ne limităm la acele genuri de inteligență care există deocamdată. De aceea definiția pe care am dat-o în primul capitol și modul în care voi folosi cuvântul în întreaga carte vizează un spectru atât de amplu:

intelența = capacitatea de a îndeplini obiective complexe

Definiția e suficient de largă pentru a le include pe toate cele sus-menționate, odată ce înțelegerea, conștiința de sine, rezolvarea problemelor, învățarea etc. sunt toate exemple de obiective complexe. Este de asemenea suficient de largă pentru a subsuma definiția din *Dicționarul Oxford* – „capacitatea de a dobândi și folosi cunoștințe și deprinderi” –, fiindcă poți avea drept obiectiv să folosești cunoștințe și deprinderi.

Pentru că există multe obiective posibile, există și multe tipuri de inteligență posibile. Conform definiției noastre, n-are sens să cuantifici inteligența oamenilor, a animalelor sau a mașinilor printr-un singur număr cum e IQ.* Care e mai inteligent, un program de calculator care joacă numai șah, sau unul care joacă numai go? Nu există un răspuns rezonabil, căci sunt două lucruri care nu pot fi comparate. Putem totuși spune că mai inteligent e un al treilea program care se dovedește a fi cel puțin la fel de bun ca ele la îndeplinirea tuturor obiectivelor, și strict mai bun în cel puțin o privință (să câștige la șah, de pildă).

Nu prea are sens nici să despici firul în patru pentru a stabili dacă ceva este sau nu inteligent în cazurile-limită, odată ce există un spectru al capacităților, nu e vorba aici de totul-sau-nimic. Ce oameni au capacitatea de a îndeplini obiectivul de a vorbi? Nou-născuții?

* Pentru a înțelege asta, închipuiți-vă cum ați reacționa dacă cineva ar vrea să cuantifice cu un singur număr, „coeficientul athletic”, capacitatea de a se descurca la probele olimpice, așa încât sportivul cu cel mai mare coeficient să câștige medaliile de aur la toate probele. (N. a.)

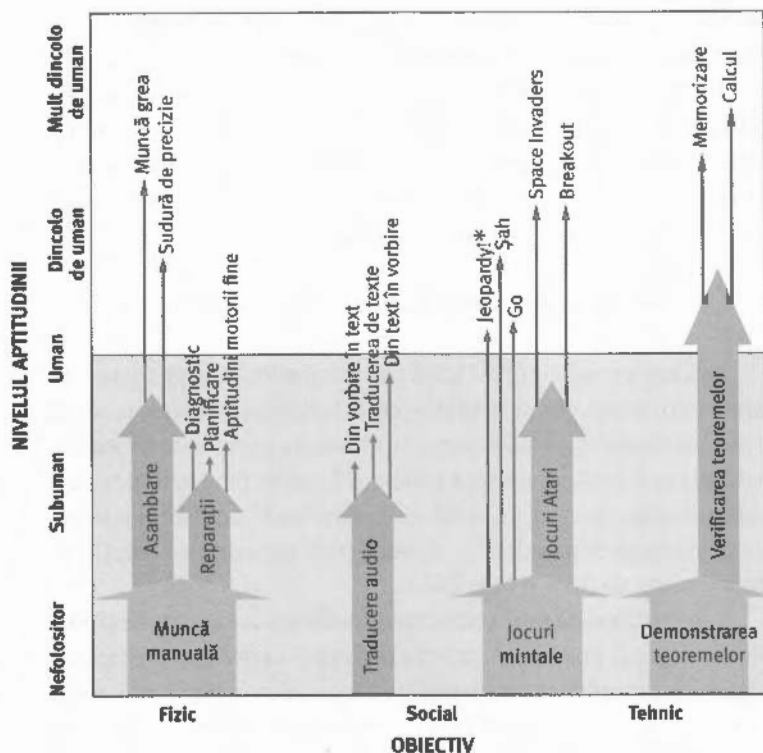


Figura 2.1 Inteligența, definită drept capacitatea de a îndeplini obiective complexe, nu poate fi măsurată printr-un singur IQ, ci doar printr-un spectru de capacități în funcție de obiective. Fiecare săgeată arată cât de apte sunt azi cele mai bune sisteme IA pentru a îndeplini diferite obiective, indicând faptul că IA de azi este *îngustă*, fiecare sistem fiind capabil să îndeplinească doar anumite obiective. Dimpotrivă, inteligența umană are un spectru remarcabil de larg: un copil sănătos poate învăța să fie superior în aproape toate privințele.

Nu. Realizatorii de emisiuni radio? Da. Cum rămâne însă cu țâncii care cunosc zece cuvinte? Sau cinci sute de cuvinte? Unde tragi linia de demarcație? Am folosit deliberat cuvântul „complex” în definiția de mai sus, fiindcă n-are rost să faci o delimitare artificială

* Jeopardy! – joc dintr-o emisiune TV din America. (N. t.)

între inteligență și non-inteligență, ci e mai util să cuantifici gradele capacității de a îndeplini diferite obiective.

Pentru a clasifica diversele inteligențe într-o taxonomie, o altă distincție importantă e cea între inteligența *îngustă* și cea *largă*. Calculatorul IBM pentru șah Deep Blue, care l-a învins pe campionul mondial Gari Kasparov în 1997, nu putea îndeplini decât foarte îngusta sarcină de a juca șah – în ciuda unui hardware și software impresionant, nu-l putea bate nici măcar pe un copil de patru ani la „x și o”. Sistemul IA DQN al lui DeepMind produs de Google poate îndeplini o gamă ceva mai largă de obiective: poate juca zeci de jocuri pe calculator Atari la nivel uman sau mai bine. Inteligența umană e însă incomparabil mai largă, având o mulțime de aptitudini. Un copil sănătos, având la dispoziție suficient timp să se antreneze, se poate descurca destul de bine nu doar în *orice* jocuri, dar și în orice limbă, sport sau meserie. Comparând inteligența de azi a oamenilor și a mașinilor, oamenii câștigă detașat la capitolul lărgimea spectrului, în timp ce mașinile ne depășesc într-un mic număr, dar în creștere, de domenii înguste, după cum se vede în figura 2.1. Marea țintă a cercetătorilor IA e să construiască „IA generală” (IAG), capabilă să îndeplinească practic orice obiectiv, inclusiv învățarea. Vom vorbi despre asta în detaliu în capitolul 4. Termenul „IAG” a fost popularizat de cercetătorii Shane Legg, Mark Gubrud și Ben Goertzel cu sensul mai precis de inteligență artificială generală la *nivel uman*: capacitatea de a îndeplini orice obiectiv cel puțin la fel de bine ca oamenii.¹ Voi folosi definiția lor, așadar, dacă nu calific explicit acronimul (scriind „IAG supraumană”, de pildă), prin „IAG” am în vedere „IAG la nivel uman”.*

Deși cuvântul „inteligență” are de regulă o conotație pozitivă, trebuie observat că îl folosesc într-un sens complet neutru: capacitatea de a îndeplini obiective complexe, indiferent dacă ele sunt

* Unii preferă „IA la nivel uman” sau „IA tare” ca sinonime pentru IAG, dar ambele sunt problematice. Chiar și un calculator de buzunar e la nivel uman într-un sens îngust. Antonimul lui „IA tare” este „IA slabă”, dar e ciudat să numești „slabe” sisteme IA precum Deep Blue, Watson și AlphaGo. (*N. a.*)

considerate bune sau rele. O persoană inteligentă poate excela deci în a-i ajuta pe oameni sau în a-i răni. Mai este de făcut o distincție subtilă în privința obiectivelor: cui aparțin? Să presupunem că viitorul tău secretar personal robot nu are obiective proprii, dar va face tot ce-i ceri, iar tu îi ceri să pregătească o cină italiană perfectă. Dacă intră pe internet și caută rețete de mâncăruri italienești, cum să ajungă la cel mai apropiat supermarket, cum să strecoare pastele etc., iar apoi cumpără ingredientele și prepară o cină gustoasă, îl vei considera pesemne inteligent, deși scopul inițial îți aparține. De fapt, el a adoptat obiectivul tău odată ce ai făcut cererea, pentru a-l diviza într-o ierarhie de sub-obiective, de la plata la magazin până la raderea parmezanului. În acest sens, comportamentul inteligent este inexorabil legat de atingerea obiectivului.

E firesc să evaluăm dificultatea sarcinilor după posibilitățile oamenilor, ca în figura 2.1, ceea ce nu mai e valabil pentru calculatoare. Pare să fie mult mai greu să înmulțești $314\,159$ cu $271\,828$ decât să recunoști un prieten într-o fotografie, dar calculatoarele ne depășeau clar la socoteli cu mult înainte să mă nasc, în vreme ce recunoașterea imaginilor la nivel uman a devenit doar de curând posibilă. Sarcini senzomotorii de bază ni se par ușoare, deși cer resurse computaționale enorme – acesta e paradoxul lui Moravec, care se explică prin faptul că în creierul nostru pentru asemenea sarcini sunt alocate mari cantități de hardware specializat, peste un sfert din creier.

Mi-am luat libertatea să ilustrez în figura 2.2 următoarea metaforă a lui Hans Moravec:

Calculatoarele sunt mașini universale, potențialul lor se extinde uniform asupra unei întinderi nesfârșite de sarcini. Pe de altă parte, potențialul uman e mare în domenii importante pentru supraviețuire, dar mic în domenii îndepărtate de aceasta. Închipuiți-vă un „peisaj al competenței umane”, cu șesuri pentru „socoteli” și „memorare papagalicească”, dealuri pentru „demonstrarea teoremelor” și „jocul de șah” și munți înalți pentru „locomotie”, „coordonare mână-ochi” și „interacție socială”.



Figura 2.2 Ilustrarea „peisajului competenței umane” al lui Hans Moravec, în care altitudinea reprezintă dificultatea pentru calculatoare, iar nivelul în creștere al apei reprezintă ce pot face calculatoarele.

Progresul performanțelor calculatoarelor este ca apa care inundă încet peisajul. Cu o jumătate de secol în urmă a înecat șesurile, înlocuind socotitorii și arhivarii, dar lăsându-ne uscați pe cei mai mulți dintre noi. Acum apa a ajuns la poalele dealurilor, iar avanposturile noastre bat în retragere. Ne simțim siguri pe crestele munților, dar în ritmul actual și ele vor fi sub ape peste încă o jumătate de secol. Propun să ne construim arce când acea zi se va apropia, și să adoptăm o viață de navigatori.²

În deceniile de când Moravec a scris aceste rânduri, nivelul mării a crescut continuu, după cum a prezis, unele dealuri (inclusiv șahul) au fost inundate, iar apa ar putea ajunge într-o bună zi la un punct critic, declanșând schimbări spectaculoase. Nivelul critic este cel în care mașinile încep să poată proiecta IA. Până atunci, creșterea nivelului mării e provocată de oameni care perfecționează mașini; după aceea, creșterea poate fi controlată de mașini care perfecționează mașini, fiind pesemne mult accelerată și inundând rapid întregul uscat. Aceasta e fascinantă și controversata idee de *singularitate*, despre care vom vorbi în capitolul 4.

Alan Turing, unul dintre pionierii informaticii, a demonstrat că dacă un calculator poate efectua un anumit set minim de operații, atunci, având la dispoziție timp și memorie suficiente, poate fi programat să facă tot ceea ce *oricare* alt calculator poate face. Mașinile care depășesc acest prag critic se numesc calculatoare universale; toate telefoanele inteligente și laptopurile de azi sunt universale în acest sens. În mod analog, cred că pragul critic de inteligență necesară pentru a proiecta IA este pragul *inteligenței universale*: având la dispoziție timp și resurse suficiente, se poate face pe sine capabilă să îndeplinească orice obiectiv la fel de bine ca *oricare* altă entitate inteligentă. De pildă, dacă se hotărăște să-și îmbunătățească aptitudinile sociale, de previziune sau de proiectare IA, o poate face. Dacă se hotărăște să descopere cum se construiește o fabrică de roboți, o poate face. Cu alte cuvinte, inteligența universală poate deveni Viața 3.0.

În ultimă instanță, inteligența ține de informație și calcul, nu de carne, sânge sau atomi de carbon – ceea ce înseamnă că nu există motive fundamentale pentru care mașinile să nu devină într-o bună zi cel puțin la fel de inteligente ca noi. Dar ce sunt de fapt informația și calculul, odată ce fizica ne spune că la nivel fundamental totul e doar materie și energie? Cum poate ceva atât de abstract, impalpabil și eteric precum informația și calculul să fie încorporat în entități fizice palpabile? În particular, cum poate manifesta comportamentul pe care-l numim inteligent o grămadă de particule mișcându-se după legile fizicii?

Dacă răspunsul ți se pare evident și consideri plauzibil ca mașinile să devină la fel de inteligente ca oamenii în acest secol – fiindcă, de pildă, ești cercetător IA –, poți sări direct la capitolul 3. În caz contrar, află că am scris următoarele trei secțiuni special pentru tine.

CE ESTE MEMORIA?

Când spunem că un atlas conține *informație* despre lume, înțelegem că există o relație între starea cărții (în particular, pozițiile anumitor molecule care dau culoare literelor și imaginilor) și starea lumii (de pildă, pozițiile continentelor). Dacă continentele s-ar afla în locuri diferite, atunci și acele molecule s-ar afla în locuri diferite. Noi, oamenii, folosim o gamă largă de dispozitive pentru stocarea informației, de la cărți și creiere la discuri dure, toate având această proprietate: starea lor poate fi pusă în corespondență cu (deci ne poate informa despre) starea altor lucruri care ne interesează.

Ce proprietate fizică fundamentală, care le face utile ca dispozitive de memorie (de stocare a informației) au ele în comun? Răspunsul este că se pot afla în stări diferite de durată lungă – suficient de lungă pentru a codifica informația până ai nevoie de ea. Ca exemplu simplu, să presupunem că așezi o bilă pe o suprafață deluroasă cu 16 văi, ca în figura 2.3. După ce bila se rostogolește și se oprește, ea se va afla într-unul dintre cele 16 locuri, așa încât poți folosi poziția ei pentru a ține minte orice număr de la 1 la 16.

Acest dispozitiv de memorie e destul de rezistent, fiindcă chiar dacă e puțin scuturat de forțe externe, bila e probabil să rămână în aceeași vale în care ai așezat-o, deci știi mai departe ce număr e stocat. Motivul pentru care această memorie e atât de stabilă este că ridicarea bilei din vale necesită mai multă energie decât e probabil să ofere perturbațiile aleatorii. Aceeași idee poate sta la baza unor memorii mult mai generale decât în cazul unei bile care se rostogolește: energia unui sistem fizic complicat poate depinde de tot felul de proprietăți fizice, chimice, electrice și magnetice, iar câtă vreme e nevoie de energie pentru a scoate sistemul din starea pe care vrei s-o memorezi, acea stare va fi stabilă. De aceea solidele au multe stări cu durată lungă de viață, dar nu și lichidele și gazele: dacă gravezi numele cuiva pe un inel de aur, informația va rămâne acolo ani de zile, fiindcă remodelarea aurului presupune multă energie, dar dacă scrii numele pe

suprafața unui lac, se pierde instantaneu, căci suprafața apei își schimbă fără efort forma.

Cel mai simplu dispozitiv de memorie are doar două stări stabile (figura 2.3). Ne putem deci închipui că el codifică un număr binar (un „bit“), adică un zero sau un unu. Informația stocată de orice dispozitiv mai complicat poate fi în mod echivalent stocată în biți multipli: de pildă, împreună, cei patru biți din figura 2.3 (dreapta) se pot afla în $2 \times 2 \times 2 \times 2 = 16$ stări diferite, adică 0000, 0001, 0010... 1111, deci au exact aceeași capacitate de memorare ca mai complicatul sistem cu 16 stări (din stânga). Putem privi așadar biții ca pe atomi de informație – bucata cea mai mică de informație care nu mai poate fi divizată și care se poate combina pentru a alcătui orice informație. De pildă, tocmai am tastat cuvântul „word“, iar laptopul meu l-a memorat ca secvența de 4 numere 119 111 114 100, stocând fiecare număr ca 8 biți (fiecare literă mică e reprezentată printr-un număr egal cu 96 plus ordinea ei în alfabeta). Când ating tasta *w*, laptopul îmi afișează pe ecran litera *w*, iar această imagine e de asemenea reprezentată prin biți: 32 de biți specifică culoarea fiecăruia dintre milioanele de pixeli ai ecranului.

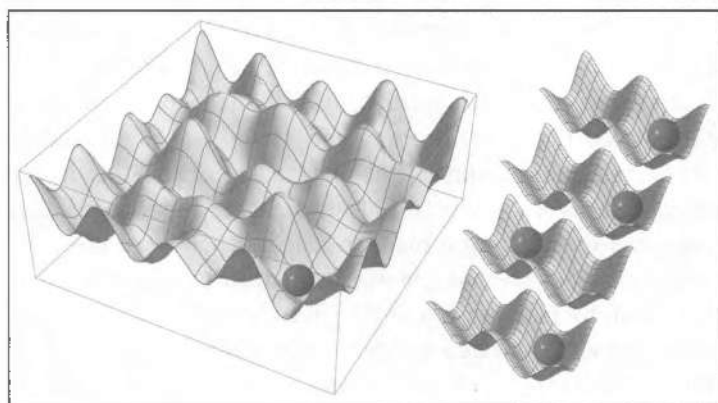


Figura 2.3 Un sistem fizic e un dispozitiv de memorie util dacă se poate afla în mai multe stări stabile. Bila din stânga poate codifica patru biți de informație desemnând în care dintre cele $2^4 = 16$ văi se află. Împreună, cele patru bile din dreapta codifică și ele patru biți de informație – fiecare câte un bit.

Din moment ce sistemele cu două stări sunt ușor de produs și de manevrat, majoritatea calculatoarelor moderne stochează informația ca biți, însă acești biți sunt realizați într-o gamă largă de moduri. Pe un DVD, fiecare bit corespunde prezenței sau absenței unei adâncituri microscopice într-un anumit loc pe suprafața de plastic. Pe un disc dur, fiecare bit corespunde unui punct de pe suprafață care e magnetizat într-unul din două moduri. În memoria de lucru a laptopului meu, fiecare bit corespunde poziției anumitor electroni, determinând dacă un dispozitiv numit microcondensator e încărcat. Unele tipuri de biți pot fi transportate, chiar și cu viteza luminii, ca într-o fibră optică ce transmite e-mailul tău, fiecare bit corespunzând unui fascicul laser mai intens sau mai slab la un moment dat.

Inginerii preferă să codifice biții în sisteme care sunt nu doar stabile și de pe care e ușor de citit (ca un inel de aur), ci pe care e și ușor de scris: modificarea stării discului tău dur necesită mult mai puțină energie decât gravarea auralui. Ei preferă de asemenea sistemele cu care e comod de lucrat și care sunt ieftine. Pe de altă parte însă, nu le pasă cum sunt reprezentați biții ca obiecte fizice – după cum nu-i pasă nici utilizatorului, fiindcă pur și simplu nu contează! Dacă-i trimiți prin e-mail prietenului tău un document să-l imprime, informația poate fi copiată într-o succesiune rapidă din magnetizările de pe discul tău dur în sarcinile electrice ale memoriei de lucru a calculatorului, undele radio din rețeaua wireless, tensiunile din router, pulsurile laser dintr-o fibră optică, și în cele din urmă în moleculele unei foi de hârtie. Cu alte cuvinte, *informația are propria ei viață, independentă de substratul ei fizic!* De regulă, ne interesează doar acest aspect al informației independent de substrat: dacă prietenul te sună să discuțați despre documentul pe care i l-ai trimis, pesemne că nu vrea să vorbească despre tensiuni sau molecule. Acesta e primul nostru indiciu privind felul în care ceva impalpabil precum inteligența poate fi încorporat în ceva palpabil, și vom vedea curând că ideea independenței de substrat e mult mai profundă, incluzând nu doar informația, ci și calculul și învățarea.

Datorită acestei independențe de substrat, inginerii au putut înlocui în repetate rânduri dispozitivele de memorie din calculatoare cu unele tot mai bune, bazate pe noi tehnologii, fără să fie nevoie de schimbări în software. Rezultatul a fost spectaculos, după cum se vede în figura 2.4: în ultimele șase decenii, prețul memoriei calculatoarelor s-a înjumătățit cam la fiecare doi ani. Discurile dure au devenit de peste 100 de milioane de ori mai ieftine, iar memoriile rapide, utile în calcule, nu doar pentru stocare, au devenit de 10 miliarde de ori mai ieftine. Dacă ai obține o asemenea reducere de 99,9999999999% la toate cumpărăturile, ai putea cumpăra toate imobilele din New York cu vreo 10 cenți, și tot aurul care a fost vreodată extras cu un dolar.

Pentru mulți dintre noi, aceste îmbunătățiri spectaculoase ale tehnologiei memoriei se leagă de povești personale. Țin minte că

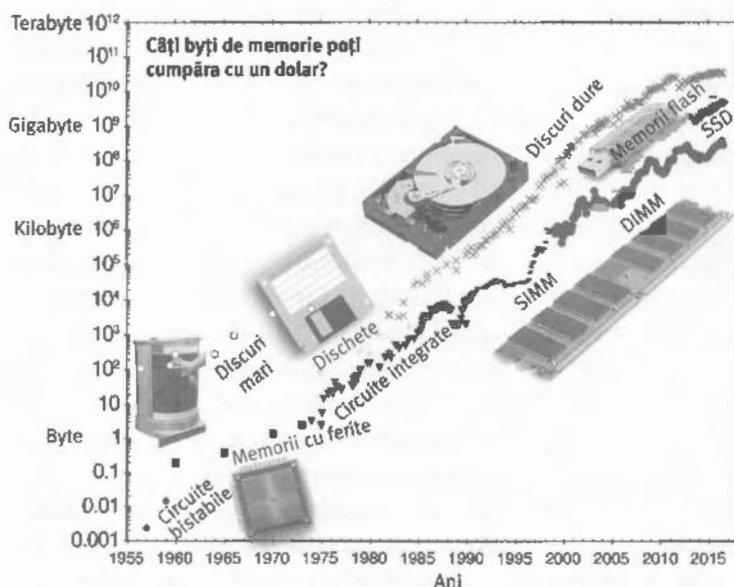


Figura 2.4 În ultimele șase decenii, prețul memoriei calculatoarelor s-a înjumătățit cam la fiecare doi ani, corespunzând unui factor de ieftinire de o mie cam la fiecare douăzeci de ani. Un byte reprezintă opt biți. Date obținute prin amabilitatea lui John McCallum din <http://www.jcmit.net/memoryprice.htm>.

în timpul liceului am muncit într-un magazin de dulciuri ca să-mi cumpăr un calculator cu o memorie de 16 kilobyți, iar atunci când am făcut și am vândut un procesor de text pentru el împreună cu colegul meu de clasă Magnus Bodin, a trebuit să-l scriem într-un cod-mașină ultracompact, ca să lăsăm suficientă memorie pentru textul care trebuia procesat. După ce m-am obișnuit cu dischetele de 70 kB, am fost uimit de dischetele mai mici, de 3,5 țoli, care stocau 1,44 MB, putând conține o carte întreagă, pentru a descoperi apoi discurile dure de 10 MB – pe care abia încapă un cântec din descărcările de azi. Aceste amintiri din adolescență mi s-au părut aproape ireale ieri, când am cumpărat cu vreo sută de dolari un disc dur cu o capacitate de 300 000 de ori mai mare.

Ce putem spune despre dispozitivele de memorie care n-au fost proiectate de oameni, ci au evoluat? Biologii nu știu deocamdată care a fost prima formă de viață care și-a copiat structura de-a lungul generațiilor, dar trebuie să fi fost foarte mică. O echipă condusă de Philipp Holliger de la Universitatea Cambridge a creat în 2016 o moleculă de ARN care codă 412 biți de informație genetică și putea copia secvențe de ARN mai lungi decât ea însăși – ceea ce venea în sprijinul ipotezei „lumii ARN”, după care viața primitivă pe Pământ a implicat scurte fragmente de ARN autoreplicatoare. Cel mai mic dispozitiv de memorie cunoscut care a evoluat este genomul bacteriei *Candidatus Carsonella ruddii*, stocând circa 40 de kilobyți, în vreme ce ADN-ul uman stochează circa 1,6 gigabyți, comparabil cu un film descărcat. După cum am spus în capitolul precedent, creierul nostru stochează mult mai multă informație decât genele noastre: electric aproximativ 10 gigabyți (specificând care dintre cei 100 de miliarde de neuroni sunt activați în orice moment) și chimic/biologic 100 de terabyți (specificând cât de puternic sunt legați diverși neuroni prin sinapse). Compararea acestor cifre cu cele ale memoriilor mașinilor arată că cele mai bune calculatoare din prezent pot depăși orice sistem biologic – la un preț care în 2016 era de câteva mii de dolari, și care scade rapid.

Memoria din creier funcționează diferit de cea din calculator, atât prin felul în care e construită, cât și prin felul în care e folosită. Dacă accesezi datele dintr-un calculator specificând unde sunt

stocate, regăsești amintirile din creier specificând *ce anume* e stocat. Fiecare grup de biți din memoria calculatorului are o adresă numerică, iar pentru a găsi un fragment de informație, calculatorul specifică la ce adresă să cauți, la fel cum ți-aș spune eu „Du-te la biblioteca mea, ia a cincea carte din dreapta de pe raftul de sus și spune-mi ce scrie la pagina 314”. În schimb, regăsești informația din creier la fel cum folosești un motor de căutare: specifice un fragment de informație sau ceva legat de ea, și informația apare. Dacă îți spun „a fi sau a nu”, sau dacă tastez asta pe Google, sunt șanse să ajung la „A fi sau a nu fi, aceasta-i întrebarea”. Metoda funcționează probabil și dacă folosesc altă parte din citat sau amestec lucrurile. Asemenea sisteme de memorie se numesc *auto-asociative*, fiindcă ne amintim prin asociație, nu prin adresă.

Într-un celebru articol din 1982, fizicianul John Hopfield arăta că o rețea de neuroni interconectați poate funcționa ca o memorie auto-asociativă. Ideea de bază mi se pare foarte frumoasă, și e valabilă pentru orice sistem fizic cu stări stabile multiple. De pildă, să considerăm o bilă pe o suprafață cu două văi, ca sistemul cu un bit din figura 2.3, și să modelăm suprafața așa încât coordonatele x ale celor două puncte de minim unde bila se află în repaus sunt $x = \sqrt{2} \approx 1,41421$ și respectiv $x = \pi \approx 3,14159$. Dacă ții minte doar că π este aproape de 3, pui pur și simplu bila la $x = 3$ și vezi cum se apropie de o valoare mai exactă a lui π rostogolindu-se spre minimumul cel mai apropiat. Hopfield a înțeles că o rețea complexă de neuroni oferă un peisaj asemănător, cu foarte multe valori minime ale energiei în care sistemul se poate stabiliza, și s-a demonstrat mai târziu că poți comprima 138 de amintiri diferite la fiecare o mie de neuroni, fără a provoca mari confuzii.

CE ESTE CALCULUL?

Am văzut că un obiect fizic poate memora informație. Dar cum poate el calcula?

Un calcul e o transformare a stării unei memorii în altă stare. Cu alte cuvinte, calculul ia informația și o transformă, punând

în aplicare ceea ce matematicienii numesc o *funcție*. Privesc funcția ca pe un fel de mașină de tocat carne pentru informație, după cum e ilustrat în figura 2.5: pui informația sus, rotești de manipulează și obții informația procesată jos – și poți repeta asta de câte ori vrei, cu diferite input-uri. Procesarea informației e deterministă, în sensul că, dacă o repeți cu același input, obții mereu același output.

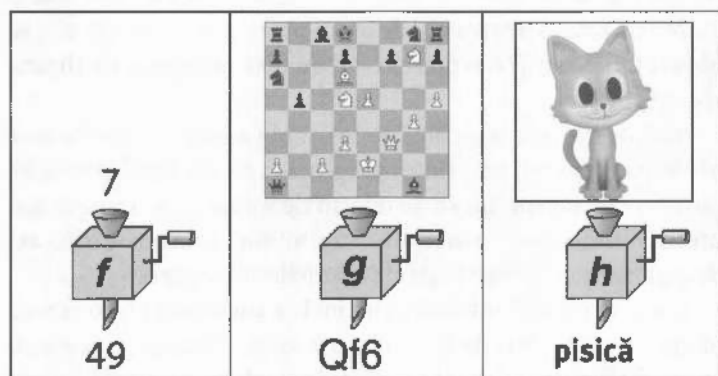


Figura 2.5 Un *calcul* ia informația și o transformă, punând în aplicare ceea ce matematicienii numesc *funcție*. Funcția *f* (din stânga) ia biți reprezentând un număr și îi ridică la pătrat. Funcția *g* (din mijloc) ia biți reprezentând o poziție de șah și calculează cea mai bună mutare pentru alb. Funcția *h* (din dreapta) ia biți reprezentând o imagine și calculează un nume care o descrie.

Deși descriptiv sună simplu, ideea de funcție e incredibil de generală. Unele funcții sunt banale, cum este cea numită *NOT*, la care input-ul e un singur bit, iar output-ul inversul lui, transformându-l astfel pe zero în unu și viceversa. Funcțiile pe care le învățăm la școală corespund de regulă butoanelor unui calculator de buzunar, introducând unul sau mai multe numere și obținând un singur număr – de pildă, funcția x^2 îți dă numărul introdus înmulțit cu el însuși. Alte funcții pot fi extrem de complicate. De exemplu, dacă dispui de o funcție la care input-ul sunt biți reprezentând o poziție de șah arbitrară, iar output-ul sunt biți reprezentând cea mai bună mutare următoare posibilă, poți să câștigi Campionatul

Mondial de Șah pe Calculator. Dacă ai o funcție la care input-ul sunt toate datele financiare din lume, iar output-ul sunt cele mai bune acțiuni de cumpărat, vei deveni repede foarte bogat. Mulți cercetători IA se ocupă în întreaga lor carieră cu punerea în aplicare a anumitor funcții. De exemplu, obiectivul cercetărilor de traducere automată este de a pune în aplicare o funcție ai căror biți de input reprezintă un text într-o limbă, iar output-ul, un text în altă limbă, în vreme ce obiectivul cercetărilor de etichetare automată este ca pentru biți de input prezentând o imagine să se obțină biți de output reprezentând un text care o descrie (figura 2.5, dreapta).

Cu alte cuvinte, dacă poți pune în aplicare funcții foarte complexe, poți construi o mașină inteligentă capabilă să îndeplinească obiective foarte complexe. Se reformulează astfel în termeni mai preciși întrebarea de la începutul capitolului: cum poate o bucată de materie neînsuflețită calcula o funcție complicată?

În loc să rămână imobilă ca un inel de aur sau alt dispozitiv de memorie static, ea trebuie să aibă o *dinamică* complexă, așa încât starea ei viitoare să depindă de starea actuală într-un mod complicat (și, e de dorit, controlabil/programabil). Aranjamentul ei atomic trebuie să fie mai puțin ordonat decât cel al unui solid rigid, în care nu se modifică nimic interesant, dar mai ordonat decât cel al unui lichid sau al unui gaz. Mai precis, vrem ca sistemul să aibă proprietatea ca, dacă îl punem într-o stare ce codifică informația de input, s-o lase un timp să evolueze conform legilor fizicii, apoi să interpreteze starea finală ca pe informația de output, iar output-ul să fie funcția dorită a input-ului. Dacă așa se întâmplă, atunci putem spune că sistemul calculează funcția noastră.

Ca prim exemplu, să examinăm cum putem construi o funcție foarte simplă (dar importantă) numită *poartă NAND** din materie neînsuflețită. Input-ul e format din doi biți, output-ul dintr-unul: output-ul e o dacă ambii biți de input sunt 1; în toate celelalte

* NAND e prescurtarea de la NOT AND („nu și”). O poartă AND dă ca output 1 dacă primul input e 1 și al doilea input e 1, deci NAND face exact pe dos. (N. a.)

cazuri, output-ul e 1. Dacă conectăm două comutatoare în serie cu o baterie și un electromagnet, prin electromagnet va circula curent doar dacă primul comutator și al doilea sunt închise. Să plasăm un al treilea comutator sub electromagnet, ca în figura 2.6, așa încât electromagnetul îl va deschide atunci când circulă curent prin el. Dacă interpretăm primele două comutatoare ca pe biții de input, iar pe al treilea ca pe bitul de output (cu 0 = comutator deschis și 1 = comutator închis), avem o poartă NAND: al treilea comutator e deschis doar dacă primele două sunt închise. Există multe alte moduri mai practice de a construi porți NAND – de pildă, folosind tranzistori (figura 2.6). În calculatoarele de azi porțile NAND sunt de regulă constituite din tranzistori microscopici și alte componente care pot fi automat gravate pe plăci de siliciu.

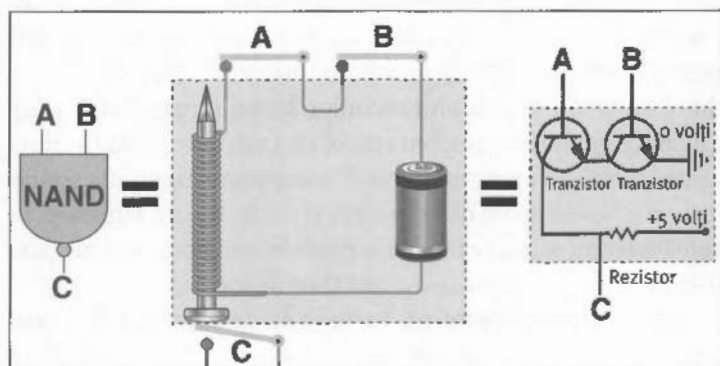


Figura 2.6 O așa-numită poartă NAND primește ca input doi biți, A și B, și calculează ca output un bit, C, după regula că $C = 0$ dacă $A = B = 1$, și $C = 1$ în celelalte cazuri. Numeroase sisteme fizice pot fi folosite ca porți NAND. În exemplul din mijloc, comutatoarele sunt interpretate ca biți (0 = deschis, 1 = închis), iar când A și B sunt închise, un electromagnet deschide comutatorul C. În exemplul din dreapta, tensiunile electrice sunt interpretate ca biți (1 = cinci volți, 0 = zero volți), iar atunci când potențialul în A și B este cinci volți, cei doi tranzistori conduc electricitatea, iar potențialul în C scade aproximativ la zero volți.

Există în informatică o remarcabilă teoremă care spune că porțile NAND sunt *universale*, în sensul că poți pune în aplicare orice funcție bine definită doar conectând porți NAND.* Dacă poți construi așadar suficiente porți NAND, poți crea un dispozitiv care să calculeze orice! Pentru a vă da o idee despre asta, în figura 2.7 arăt cum poți înmulți numere folosind numai porți NAND.

Cercetătorii Norman Margolus și Tommaso Toffoli de la MIT au creat termenul *computroniu* pentru a desemna orice substanță care poate efectua calcule. Singura condiție pentru computroniu este, după cum am văzut, să poată pune în funcțiune porți NAND conectate oricum ne dorim, și există o mulțime de tipuri de computroniu. O altă variantă simplă este de a înlocui porțile NAND cu porți NOR, care dau ca output 1 doar dacă ambele input-uri sunt 0. În secțiunea următoare vom vorbi despre rețelele neuronale, care pot de asemenea efectua calcule, deci acționează ca un computroniu. Stephen Wolfram a arătat că același lucru e valabil pentru dispozitive simple numite automate celulare, care își modifică biții în funcție de biții vecinilor. Încă din 1936 Alan Turing a demonstrat într-un celebru articol că o mașină simplă (numită acum „mașină Turing universală”) care poate manipula simboluri pe o bandă poate de asemenea efectua calcule arbitrare. În concluzie, nu numai că materia poate efectua orice calcul bine definit, dar o poate face într-o mulțime de moduri.

După cum am menționat, Turing a demonstrat în acel articol din 1936 și ceva mai profund: dacă un calculator poate efectua un anumit set de operații simple, atunci e *universal*, în sensul că, având suficiente resurse, poate face tot ce face orice alt calculator. A demonstrat că mașina Turing e universală, iar noi tocmai am văzut că această familie de calculatoare universale include obiecte diverse precum o rețea de porți NAND și o rețea de neuroni

* Prin „funcție bine definită” înțeleg ceea ce matematicienii și informaticienii numesc „funcție calculabilă” — o funcție care poate fi calculată cu un computer ipotetic, având memorie și timp nelimitate. Alan Turing și Alonzo Church au demonstrat că există funcții care pot fi descrise, dar nu sunt calculabile. (N. a.)

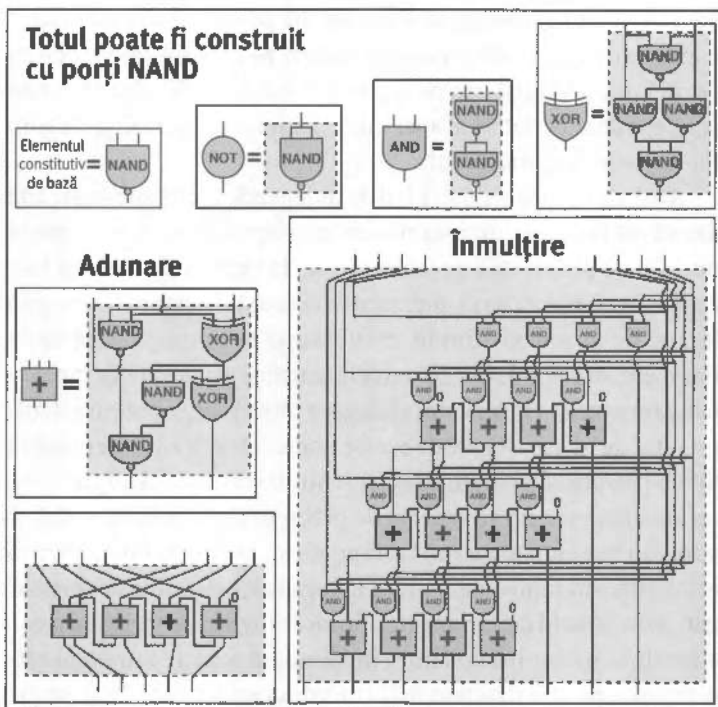


Figura 2.7 *Orice* calcul bine definit poate fi efectuat combinând ingenios doar porți NAND. De pildă, modulele de adunare și înmulțire de mai sus primesc ca input două numere binare reprezentate prin 4 biți și dau ca output un număr binar reprezentat prin 5 biți și, respectiv, 8 biți. Modulele mai mici NOT, AND, XOR și + (care adună trei biți într-un număr binar de 2 biți) sunt, la rândul lor, construite cu porți NAND. Înțelegerea deplină a acestei figuri e dificilă și inutilă pentru lectura cărții de față. Am inclus-o aici doar ca să ilustrez ideea de universalitate – și din pedanterie.

interconectați. Stephen Wolfram a arătat că *majoritatea* sistemelor fizice nebanale, de la sisteme climatice la creiere, ar putea fi calculatoare universale dacă ar avea dimensiuni arbitrare de mari și un timp de viață lung.

Faptul că exact același calcul poate fi efectuat pe *orice* calculator universal arată că într-adevăr *calculul e independent de substrat* – la fel ca informația, are propria lui viață. Dacă ai fi un personaj

conștient superinteligent dintr-un joc pe calculator din viitor, n-ai avea cum să-ți dai seama dacă rulezi pe un desktop Windows, pe un laptop Mac OS sau pe un telefon Android, fiindcă ai fi independent de substrat. N-ai avea cum să știi nici ce tip de tranzistori folosește microprocesorul.

Am înțeles importanța ideii de independență de substrat când am văzut formele interesante sub care apare în fizică. De pildă, unele: au proprietăți precum viteza, lungimea de undă și frecvența – iar noi, fizicienii, putem studia ecuațiile cărora li se supun fără să fie nevoie să știm în ce substanță se propagă acele unde. Când auzim ceva, detectăm undele acustice provocate de moleculele care se mișcă în amestecul de gaze numit aer, și putem calcula tot felul de lucruri legate de aceste unde – cum scade intensitatea lor cu pătratul distanței, cum își schimbă direcția când trec prin uși deschise și cum se reflectă de pereți generând ecoul – fără să știm din ce e alcătuit aerul. De fapt, nici n-avem nevoie să știm că e alcătuit din molecule, putem ignora toate detaliile despre oxigen, azot, dioxid de carbon etc., fiindcă singura proprietate a substratului care contează și intră în faimoasa ecuație a undelor este un număr pe care îl putem măsura: viteza undei, care, în acest caz, este de aproximativ 300 de metri pe secundă. Această ecuație a undelor, pe care am predat-o primăvara trecută studenților mei de la MIT, a fost descoperită și folosită cu mult înainte ca fizicienii să știe de existența atomilor și moleculelor!

Acest exemplu ilustrează trei aspecte importante. În primul rând, independența de substrat nu înseamnă că substratul nu-i necesar, ci că majoritatea detaliilor lui nu contează. Evident, nu poți avea unde acustice într-un gaz dacă nu există gazul, dar e suficient să existe un gaz oarecare. În mod asemănător, nu poți avea calcul fără materie, dar orice materie e bună, cu condiția să poată fi aranjată în porți NAND, să poată conecta neuroni sau alte elemente constitutive de bază care permit calculul universal. În al doilea rând, fenomenul independenței de substrat își are propria lui viață. O undă poate străbate un lac chiar dacă nici una dintre moleculele de apă n-o face – ele doar urcă și coboară, la fel

ca „valurile“ din tribunele stadioanelor. În al treilea rând, adesea ne interesează numai aspectul independent de substrat: pe un surfer îl preocupă poziția și înălțimea valului, nu și compoziția lui moleculară. Am văzut că acest lucru e valabil în cazul informației, iar el este valabil și pentru calcul: dacă doi programatori încearcă să găsească eroarea din codul lor, probabil că nu discută despre tranzistori.

Am ajuns acum să răspundem la întrebarea cum se face că ceva fizic palpabil poate da naștere la ceva ce pare atât de impalpabil, abstract și eteric precum inteligența: pare non-fizic pentru că e independent de substrat, are propria lui viață care nu depinde de detaliile fizice și nu le reflectă. Pe scurt, calculul e un tipar în aranjamentul spațio-temporar al particulelor, iar ceea ce contează e tiparul, nu particulele.

Cu alte cuvinte, hardware-ul e materia, iar software-ul e tiparul. Această independență de substrat a calculului arată că IA e posibilă: inteligența n-are nevoie de carne, sânge sau atomi de carbon.

Datorită acestei independențe de substrat, ingineri destoinici au putut înlocui în repetate rânduri tehnologiile din computerele noastre cu unele mult mai bune, fără a schimba software-ul. Rezultatul a fost la fel de spectaculos ca în cazul dispozitivelor de memorie. Așa cum se vede în figura 2.8, prețul calculului s-a tot înjumătățit cam la fiecare doi ani timp de peste un secol, ajungând să fie de un milion de milioane de milioane (10^{18}) de ori mai ieftin decât pe vremea când s-au născut bunicii mei. Dacă totul s-ar ieftini de un milion de milioane de milioane de ori, atunci ai putea cumpăra cu o sutime de cent toate bunurile și serviciile produse pe Pământ anul acesta. Scăderea spectaculoasă a costurilor este desigur motivul principal pentru care calculul a pătruns acum peste tot, trecând de la centrele de calcul de dimensiunea unei clădiri de mai ieri în casele, mașinile și buzunarele noastre – ajungând până și în locuri neașteptate cum e încălțămînta sport.

De ce tehnologia noastră își dublează puterea la intervale regulate, prezentând ceea ce matematicienii numesc creștere exponențială? De ce se întâmplă așa nu doar cu miniaturizarea

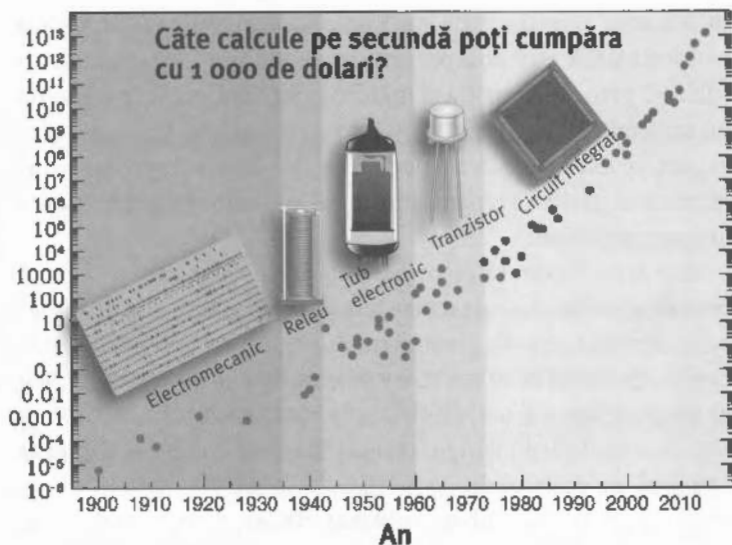


Figura 2.8 Începând din 1900, prețul calculului s-a tot înjumătățit cam la fiecare doi ani. Acest grafic prezintă puterea de calcul măsurată în operații în virgulă mobilă pe secundă (*floating-point operations per second* – FLOPS) care poate fi cumpărată cu 1 000 de dolari.³ Calculul care definește o operație în virgulă mobilă corespunde cu circa 105 operații logice elementare precum inversarea unui bit sau o evaluare NAND.

tranzistorilor (tendință cunoscută drept *legea lui Moore*), dar și cu calculul în ansamblul lui (figura 2.8), cu memoria (figura 2.4) și cu o mulțime de alte tehnologii, de la secvențializarea genomului la imagistica cerebrală? Ray Kurzweil numește acest fenomen de dublare constantă „legea beneficiilor accelerate”.

Toate exemplele de dublare constantă pe care le cunosc din natură au aceeași cauză fundamentală, iar cel tehnologic nu face excepție: fiecare pas îl creează pe următorul. De pildă, tu însuși ai cunoscut o creștere exponențială imediat ce ai fost conceput: fiecare celulă s-a divizat, dând naștere zilnic cam la două celule și făcând ca numărul tău total de celule să crească zi de zi ca 1, 2, 4, 8, 16 și așa mai departe. Conform celei mai plauzibile teorii științifice privind originea noastră cosmică, numită *inflație*, universul

nostru a crescut la început la fel ca tine, dublându-și în mod repetat dimensiunea la intervale regulate, până când un grăunte mult mai mic și mai ușor decât un atom a devenit mai mare decât toate galaxiile pe care le-am văzut vreodată cu telescoapele noastre. Cauza a fost din nou un proces în care fiecare pas l-a declanșat pe următorul. La fel progresează și tehnologia: odată ce tehnologia devine de două ori mai puternică, ea poate fi adesea folosită pentru a proiecta și construi o tehnologie de două ori mai puternică, dublarea continuând în spiritul legii lui Moore.

Dar ceva care se dublează regulat, precum puterea tehnologiei noastre, ne-am putea gândi că trebuie să se oprească la un moment dat. Desigur, legea lui Moore nu se va mai aplica, fiindcă există o limită fizică în miniaturizarea tranzistorilor. Unii presupun în mod greșit că legea lui Moore e sinonimă cu dublarea constantă a puterii noastre tehnologice. Ray Kurzweil arată însă că legea lui Moore implică nu prima, ci a cincea paradigmă tehnologică pentru a provoca creșterea exponențială a calculului din figura 2.8: de fiecare dată când tehnologia a încetat să se perfecționeze, ea a fost înlocuită de una mai bună. Când n-am mai putut micșora tuburile electronice, le-am înlocuit cu tranzistori, iar apoi cu circuite integrate, în care electronii se deplasează în două dimensiuni. Când această tehnologie își atinge limitele, există multe alte variante pe care le putem încerca – de pildă, să folosim circuite tridimensionale și să punem la lucru altceva decât electronii.

Nimeni nu știe cu certitudine care va fi următorul substrat computațional, dar știm că suntem încă departe de limitele impuse de legile fizicii. Seth Lloyd, colegul meu de la MIT, a calculat care e această limită fundamentală și, după cum vom vedea în detaliu în capitolul 6, ea se află cu 33 de ordine de mărime (10^{33}) dincolo de ceea ce poate face azi o bucată de materie care calculează. Așa încât, chiar dacă dublăm mai departe puterea calculatoarelor la fiecare doi ani, vor mai trece două secole până să ajungem la frontiera ultimă.

Deși toate calculatoarele universale pot efectua aceleași calcule, unele sunt mai eficiente decât altele. De pildă, un calcul care

presupune milioane de înmulțiri nu necesită milioane de module separate de înmulțire construite din tranzistori diferiți, ca în figura 2.6, ci doar unul singur, odată ce poate fi folosit succesiv de mai multe ori cu input-urile potrivite. Pentru eficiență, majoritatea computerelor moderne se bazează pe o paradigmă în care calculele sunt divizate în pași temporali în cursul cărora informația e plimbată între modulele de memorie și modulele de calcul. Această arhitectură computațională a fost elaborată între 1935 și 1945 de Alan Turing, Konrad Zuse, John Mauchly și John von Neumann. Mai exact, memoria calculatorului stochează atât datele, cât și software-ul (un program, adică o listă de instrucțiuni pentru ce are de făcut cu datele). La fiecare pas, o unitate centrală de procesare (CPU) execută următoarea instrucțiune din program, care specifică aplicarea unei funcții simple asupra anumitor date. Partea din calculator care urmărește ce e de făcut în continuare e doar o altă zonă din memoria lui, numită *contorul de program*, care stochează numărul liniei curente din program. Pentru a sări la altă linie din program, copiezi pur și simplu numărul liniei în contorul de program – acesta e felul în care așa-numitele instrucțiuni și bucle „if” sunt executate.

Calculatoarele de azi câștigă în viteză prin *procesare paralelă*, care reutilizează ingenios modulele: dacă un calcul poate fi desfășurat în părți ce pot fi efectuate în paralel (fiindcă input-ul unei părți nu presupune output-ul alteia), atunci părțile pot fi calculate simultan de diferite componente ale hardware-ului.

Calculatorul paralel ultim este *computerul cuantic*. David Deutsch, unul dintre pionierii calcului cuantic, susține controversata idee conform căreia „computerele cuantice împart informația cu un număr uriaș de versiuni ale lor însele din multivers” și pot căpăta răspunsuri mai repede aici, în universul nostru, primind într-un anumit sens ajutor de la aceste alte versiuni.⁴ Nu știm încă dacă un computer cuantic competitiv comercial poate fi construit în următoarele decenii – asta depinde de corectitudinea felului în care înțelegem că funcționează mecanica cuantică și de capacitatea noastră de a depăși imense dificultăți tehnice –, dar companii

și guverne din întreaga lume pariază anual zeci de milioane de dolari pe această posibilitate. Deși computerele cuantice nu pot accelera calculele obișnuite, au fost elaborați algoritmi ingenioși care pot accelera anumite tipuri de calcule cum ar fi spargerea codurilor și antrenarea rețelelor neuronale. Un computer cuantic ar putea de asemenea simula eficient comportamentul sistemelor cuantice, între care atomi, molecule și materiale noi, înlocuind măsurătorile din laboratoarele de chimie, așa cum simulările pe computere tradiționale au înlocuit măsurătorile din tunelurile aerodinamice.

CE ESTE ÎNVĂȚAREA?

Deși un calculator de buzunar mă poate surclasa la un concurs de socoteli, el nu-și va îmbunătăți niciodată viteza sau precizia, oricât ar exersa. El nu învață: de pildă, de fiecare dată când apăs butonul „radical“, calculează exact aceeași funcție în exact același fel. În mod asemănător, primul program de calculator care m-a bătut la șah n-a învățat niciodată din greșelile lui, ci pur și simplu a executat funcția pe care un programator ingenios o proiectase pentru a calcula o mutare bună. Dimpotrivă, când Magnus Carlsen a pierdut primul joc de șah la vârsta de cinci ani, a început un proces de învățare care a făcut din el, optsprezece ani mai târziu, campion mondial la șah.

Capacitatea de a învăța este pesemne cel mai fascinant aspect al inteligenței generale. Am văzut deja cum o bucată de materie neînsuflită poate memora și calcula, dar cum poate ea învăța? Am văzut că găsirea răspunsului la o întrebare dificilă e echivalentă cu calculul unei funcții, și că materia aranjată corespunzător poate calcula orice funcție calculabilă. Când noi, oamenii, am construit primele calculatoare de buzunar și primele programe de șah, noi am făcut aranjarea. Pentru ca materia să învețe, trebuie să se rearanjeze pe ea însăși ca să se perfecționeze tot mai mult în calculul funcției dorite – supunându-se pur și simplu legilor fizicii.

Pentru a demistifica procesul învățării, să vedem cum sisteme fizice foarte simple pot învăța cifrele lui π și ale altor numere. Am văzut mai sus că o suprafață cu mai multe văi (figura 2.3) poate fi folosită ca dispozitiv de memorie: de pildă, dacă fundul uneia dintre văi se afla la poziția $x = \pi \approx 3,14159$ și nu există alte văi în apropiere, atunci poți pune bila la $x = 3$ și vei vedea cum sistemul calculează zecimalele lipsă lăsând bila să se rostogolească. Să presupunem acum că suprafața e alcătuită dintr-un lut moale, iar la început e perfect plată. Dacă vreun pasionat de matematică așază în mod repetat bila în pozițiile numerelor lui preferate, treptat gravitația va crea văi în acele locuri, după care suprafața de lut poate fi folosită pentru a regăsi aceste amintiri stocate. Cu alte cuvinte, suprafața de lut a învățat să calculeze cifrele unor numere precum π .

Alte sisteme fizice, cum sunt creierele, pot învăța mult mai eficient pe baza aceleiași idei. John Hopfield a arătat că rețeaua lui de neuroni interconectați menționată mai sus poate învăța într-un mod asemănător: dacă o pui în repetate rânduri în anumite stări, treptat va învăța aceste stări și se va întoarce la ele din orice stare învecinată. Dacă i-ai văzut pe toți membrii familiei tale de multe ori, atunci amintirea felului în care arată poate fi declanșată de orice se leagă de ei.

Rețelele neuronale au transformat acum atât biologia, cât și inteligența artificială, iar recent au început să domine un subdomeniu IA numit *învățarea automată* (studiul algoritmilor care se perfecționează prin experiență). Înainte de a înțelege cum pot învăța asemenea rețele, să vedem cum pot ele calcula. O rețea neuronală e pur și simplu un grup de neuroni interconectați care își pot influența reciproc comportamentul. Creierul conține vreo sută de miliarde de neuroni, cam câte stele sunt în galaxia noastră. În medie, fiecare neuron e conectat cu aproximativ o mie de alți neuroni prin legături numite *sinapse*, iar țăriile acestor circa o sută de bilioane de conexiuni sinaptice este cea care codează majoritatea informației din creier.

Putem desena schematic o rețea neuronală ca pe un ansamblu de puncte reprezentând neuronii, legate prin linii reprezentând

sinapsele (figura 2.9). În realitate, neuronii sunt dispozitive electrochimice foarte complicate, constituite din elemente cum sunt axonii și dendritele. Există diferite tipuri de neuroni, care acționează în moduri diferite, iar felul în care activitatea unui neuron influențează alți neuroni și timpul de reacție nu sunt încă pe deplin lămurite. Cercetătorii IA au arătat totuși că rețelele neuronale pot atinge performanțe de nivel uman în multe sarcini complexe chiar dacă ignorăm toate aceste detalii și înlocuim neuronii biologici reali cu unii simulați extrem de simpli, toți identici și supuși unor reguli foarte simple. Modelul curent cel mai des folosit pentru o asemenea *rețea neuronală artificială* reprezintă starea unui neuron printr-un singur număr, iar tăria fiecărei sinapse tot printr-un singur număr. În acest model, fiecare neuron își actualizează starea la intervale de timp regulate prin medierea input-urilor de la toți neuronii cu care e conectat, ponderându-le cu tăriile sinaptice, adăugând eventual o constantă, iar apoi aplicând rezultatului ceea ce se numește o *funcție de activare* pentru a calcula noua stare a neuronului.* Cel mai simplu mod de a folosi o rețea neuronală ca pe o funcție este s-o alimentăm cu informație dintr-o singură direcție, ca în figura 2.9, introducând input-ul funcției în neuronii din stratul de sus și extrăgând output-ul de la neuronii din stratul de jos.

Succesul acestor rețele neuronale artificiale simple este încă o dovadă a independenței de substrat: rețelele neuronale se pare că au o mare putere de calcul independent de detaliile construcției lor la nivel elementar. George Cybenko, Kurt Hornik, Maxwell Stinchcombe și Halbert White au demonstrat în 1989 ceva

* Dacă vă place matematica, două alegeri tipice pentru această funcție de activare sunt funcția sigmoidă $\sigma(x) \equiv 1/(1 + e^{-x})$ și funcția rampă $\sigma(x) = \max[0, x]$, dar s-a demonstrat că aproape orice funcție e bună, cu condiția să nu fie liniară (o dreaptă). Celebrul model al lui Hopfield folosește $\sigma(x) = -1$ dacă $x < 0$ și $\sigma(x) = 1$ dacă $x \geq 0$. Dacă stările neuronului sunt stocate într-un vector, atunci rețeaua e actualizată pur și simplu înmulțind vectorul cu o matrice care stochează cuplajele sinaptice, iar apoi aplică funcția $\sigma(x)$ tuturor elementelor. (N. a.)

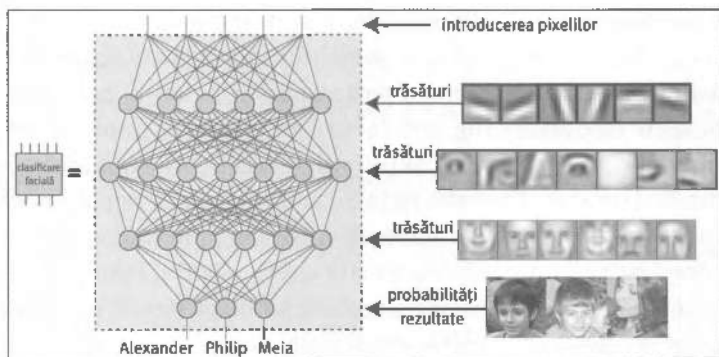


Figura 2.9 O rețea de neuroni poate calcula funcții la fel ca o rețea de porți NAND. De exemplu, rețele neuronale artificiale au fost antrenate ca, dacă introduci numere reprezentând strălucirea diferiților pixeli dintr-o imagine, să obții numere reprezentând probabilitatea ca imaginea să înfățișeze diverși oameni. Aici, fiecare neuron artificial (cerculeț) calculează o sumă ponderată a numerelor trimise către el prin conexiuni (linii) de sus, aplică o funcție simplă și transmite rezultatul în jos, fiecare strat succesiv calculând trăsături de nivel mai înalt. Rețelele tipice de recunoaștere facială conțin sute de mii de neuroni, dar în figură apar doar câțiva pentru a simplifica descrierea.

remarcabil: asemenea rețele neuronale simple sunt universale în sensul că pot calcula *orice* funcție cu o precizie oricât de mare prin ajustarea corespunzătoare a acelor numere care dau tăria sinapsei. Cu alte cuvinte, probabil că evoluția n-a făcut neuronii noștri biologici atât de complicați pentru că era necesar, ci pentru că era mai eficient – și pentru că evoluția, spre deosebire de inginerii umani, nu răsplătește proiectele simple și ușor de înțeles.

Când am aflat asta, m-a mirat cum ceva atât de simplu poate calcula ceva arbitrar de complicat. De pildă, cum poți calcula până și ceva atât de banal ca o înmulțire, când tot ce ți se permite este să calculezi sume ponderate și să aplici o singură funcție fixată? Dacă vrei să vă faceți o idee, figura 2.10 prezintă felul în care cinci neuroni pot înmulți două numere arbitrare, iar un singur neuron poate înmulți trei biți.

Deși se poate demonstra că *în teorie* poți calcula orice cu o rețea neuronală arbitrar de mare, demonstrația nu spune dacă o poți

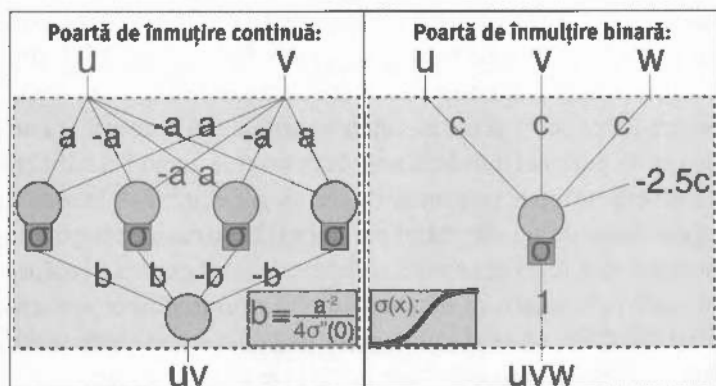


Figura 2.10 Cum poate calcula materia nu folosind porți NAND ca în figura 2.7, ci neuroni. Fără a intra în detalii, ideea esențială este că neuronii (artificiali sau biologici) pot face calcule, iar pentru înmulțire e nevoie de mai puțini neuroni decât de porți NAND. *Detalii opționale pentru pasionații de matematică:* Cerculețele efectuează adunări, pătratele aplică funcția σ , iar liniile înmulțesc cu constantele marcate pe ele. Input-urile sunt numere reale (stânga) și biți (dreapta). Înmulțirea devine arbitrar de precisă când $a \rightarrow 0$ (stânga) și $c \rightarrow \infty$ (dreapta). Rețeaua din stânga funcționează pentru orice funcție $\sigma(x)$ care e curbată în origine (cu derivata a doua $\sigma''(0) \neq 0$), ceea ce se poate demonstra prin dezvoltarea Taylor a lui $\sigma(x)$. Rețeaua din dreapta presupune ca funcția $\sigma(x)$ să se apropie de 0 și 1 atunci când x devine foarte mic, respectiv foarte mare, ceea ce se vede observând că $uvw = 1$ doar dacă $u + v + w = 3$. (Exemplele sunt dintr-un articol pe care l-am scris împreună cu studenții mei Henry Lin și David Rolnik, „Why Does Deep and Cheap Learning Work So Well?“, care poate fi găsit la <http://arxiv.org/abs/1608.08225>.) Combinând numeroase înmulțiri și adunări, poți calcula orice polinoame, despre care se știe că pot aproxima orice funcție netedă.

face în practică, având o rețea de dimensiuni rezonabile. Cu cât mă gândesc mai mult, cu atât sunt mai uimit că rețelele neuronale funcționează atât de bine.

De exemplu, să presupunem că vrem să clasificăm imagini de un megapixel în nuanțe de gri în două categorii, pisici și câini. Dacă fiecare pixel din acel milion poate lua, să zicem, 256 de valori, atunci există $256^{1000000}$ imagini posibile, iar pentru fiecare vrem să calculăm probabilitatea de a reprezenta o pisică. Asta

înseamnă că o funcție arbitrară care are ca input o imagine și ca output o probabilitate e definită printr-o listă de $256^{1000000}$ probabilități, adică mai multe numere decât atomii din universul nostru (circa 10^{78}). Și totuși, rețele neuronale cu doar mii sau milioane de parametri îndeplinesc bine sarcina acestei clasificări. Cum reușesc rețele neuronale de succes să fie „ieftine“, în sensul că au nevoie de atât de puțini parametri? La urma urmei, poți demonstra că o rețea neuronală suficient de mică pentru a încăpea în universul nostru va eșua lamentabil să aproximeze aproape toate funcțiile, reușind să îndeplinească doar o parte infimă din toate sarcinile computaționale pe care i le-ai putea atribui.

M-am gândit mult la asta și la probleme înrudite împreună cu studentul meu Henry Lin. Spre norocul meu, am avut ocazia să colaborez cu studenți minunați, iar Henry este unul dintre ei. Când a intrat prima dată în biroul meu și m-a întrebat dacă vreau să lucrez cu el, mi-am zis că mai curând eu ar trebui să-l întreb dacă îl interesează să lucreze cu mine. Acest băiat modest, prietenos, cu ochi strălucitori, din Shreveport, Louisiana, scrisese deja opt articole științifice, câștigase Premiul Forbes „30-Under-30“ și ținuse o conferință TED cu peste un milion de vizualizări – și avea doar douăzeci de ani! Un an mai târziu, am scris împreună un articol cu o concluzie surprinzătoare: matematica nu poate răspunde singură la întrebarea de ce funcționează atât de bine rețelele neuronale, este nevoie și de fizică. Am descoperit că acea clasă de funcții pe care ni le oferă fizica și asupra căreia ne trezește interesul pentru a le calcula este de asemenea o clasă extrem de redusă pentru că, din motive pe care încă nu le înțelegem pe deplin, legile fizicii sunt remarcabil de simple. Mai mult, infima parte a funcțiilor pe care rețelele neuronale le pot calcula e foarte asemănătoare cu infima parte asupra căreia fizica ne trezește interesul! Am arătat apoi că rețelele neuronale de învățare profunde (sunt numite „profunde“ dacă au în compoziție multe straturi) sunt mult mai eficiente decât cele subțiri pentru multe dintre aceste funcții de interes. De pildă, împreună cu David Rolnik, un alt student excepțional de la MIT, am demonstrat că sarcina simplă

de a înmulți n numere necesită 2^n neuroni într-o rețea cu un singur strat, dar numai circa $4n$ neuroni într-o rețea profundă. Asta explică nu doar de ce rețelele neuronale sunt la modă printre cercetătorii IA, ci și de ce am dezvoltat în creierele noastre rețele neuronale: dacă am dezvoltat creiere pentru a prezice viitorul, atunci e logic să fi dezvoltat o arhitectură computațională adaptată tocmai acelor probleme de calcul care contează în lumea fizică.

După ce am văzut cum funcționează și calculează rețelele neuronale, să ne întoarcem la problema învățării. Mai precis, cum își poate îmbunătăți capacitatea de calcul o rețea neuronală actualizându-și sinapsele?

În celebra sa carte din 1949 *The Organization of Behavior: A Neuropsychological Theory*, psihologul canadian Donald Hebb susține că dacă doi neuroni învecinați se activează mereu în același timp, cuplajul lor sinaptic se întărește, așa încât ei învață să se activeze reciproc – idee surprinsă de expresia „Fire together, wire together” („Activați împreună, cablați împreună”). Deși detaliile felului în care învață creierul sunt încă departe de a fi înțelese, iar cercetările arată că în numeroase cazuri răspunsurile sunt mult mai complicate, s-a demonstrat că până și această regulă simplă de învățare (numită învățare hebbiană) permite rețelelor neuronale să învețe lucruri interesante. John Hopfield a arătat că învățarea hebbiană permite rețelei lui neuronale ultrasimplificate să stocheze multe amintiri complexe doar prin expunerea repetată la acestea. Asemenea expunere la informație pentru a învăța din ea se numește de regulă „antrenament” când se referă la rețele neuronale artificiale (ori la animale sau oameni care capătă aptitudini), deși mai nimerit ar fi „studiu”, „educație” sau „experiență”. Rețelele neuronale artificiale din sistemele IA de azi tind să înlocuiască învățarea hebbiană cu reguli de învățare mai sofisticate, purtând denumiri întortocheate precum „retropropagare” („*back-propagation*”) și „coborâre stocastică în gradient” („*stochastic gradient descent*”), însă ideea de bază e aceeași: există o regulă deterministă simplă, asemănătoare unei legi a fizicii, prin care sinapsele sunt actualizate din timp în timp. Și, ca prin magie, această regulă

simplă poate face ca rețelele neuronale să învețe calcule extrem de complexe dacă antrenamentul se desfășoară cu cantități mari de date. Nu știm încă exact ce reguli de învățare folosește creierul nostru, dar indiferent care ar fi răspunsul, nu avem vreun indiciu că ele ar încălca legile fizicii.

După cum majoritatea computerelor digitale câștigă în eficiență împărțindu-și munca în mai mulți pași și refolosind repetat modulele de calcul, la fel fac și multe rețele neuronale artificiale și biologice. Creierul conține părți pe care informaticienii le numesc rețele *recurente* (spre deosebire de rețelele *feedforward*), în care informația poate curge în mai multe direcții (nu într-una singură), așa încât output-ul curent poate deveni apoi input. Rețeaua de porți logice din microprocesorul unui laptop este de asemenea recurentă, în următorul sens: își refolosește constant informația din trecut și lasă noua informație de input (de la tastatură, trackpad, cameră video etc.) să influențeze calculul curent, care, la rândul lui, determină informația de output către ecran, difuzor, imprimantă sau rețea wireless. În mod asemănător, rețeaua de neuroni din creier e recurentă, lăsând informația de input provenind de la ochi, urechi și alte organe senzoriale să influențeze calculul curent, care, la rândul lui, determină informația de output către mușchi.

Istoria învățării e cel puțin la fel de lungă ca însăși istoria vieții, căci orice organism autoreproducător copiază și procesează informația – comportament care a fost cumva învățat. În epoca Vieții 1.0 însă, organismele n-au învățat în cursul vieții lor: regulile de procesare a informației și de reacție erau determinate de ADN-ul lor moștenit, așa încât învățarea avea loc lent, la nivelul speciei, prin evoluție darwinistă de-a lungul generațiilor.

Cu aproximativ o jumătate de miliard de ani în urmă, anumite structuri genetice au găsit o modalitate de a face ca animale conținând rețele neuronale să poată învăța comportamente din experiențele trăite în cursul vieții. Apăruse Viața 2.0, și datorită capacității ei de a învăța incomparabil mai repede, surclasându-și competitorii, s-a răspândit fulgerător pe glob. După cum am văzut în capitoul 1, viața a progresat în privința învățării într-un ritm

tot mai rapid. O specie semănând cu maimuțele antropoide a căpătat un creier atât de bine adaptat la dobândirea cunoștințelor, încât a învățat să folosească unelte, să aprindă focul, să vorbească și să creeze o societate globală complexă. Societatea însăși poate fi privită ca un sistem care memorează, calculează și învață, toate acestea într-un ritm accelerat, căci o invenție o face posibilă pe următoarea: scris, presă tipărită, știință modernă, calculatoare, internet etc. Ce vor nota istoricii științei din viitor în continuarea listei invențiilor? Inteligență artificială, cred eu.

Progresele explozive ale memoriei și puterii de calcul ale computerelor (figurile 2.4 și 2.8) au determinat un progres spectaculos al IA, dar a trecut mult timp până ca *învățarea* automată să se maturizeze. Când calculatorul IBM Deep Blue l-a învins la șah pe campionul mondial Gari Kasparov, în 1997, atuurile lui țineau de memorie și calcul, nu de învățare. Inteligența lui computațională fusese creată de o echipă de oameni, iar Deep Blue își învingea creatorii grație vitezei lui mai mari de calcul, care îi permitea să analizeze mai multe posibilități. Când calculatorul IBM Watson l-a învins pe campionul emisiunii-concurs *Jeopardy!*, nici acesta nu s-a bazat pe învățare, ci pe aptitudini specializate programate și pe memorie și viteză superioare. Același lucru e valabil pentru majoritatea primelor realizări din robotică, de la locomotia cu picioare la automobile care se conduc singure și rachete care aterizează singure.

Dimpotrivă, forța motrice a celor mai recente realizări IA a fost *învățarea* automată. Priviți, de exemplu, figura 2.11. E ușor pentru noi să spunem ce reprezintă fotografia, dar ca să programezi o funcție pentru care input-ul să fie doar culorile tuturor pixelilor unei imagini, iar output-ul să fie o descriere precisă cum ar fi „Un grup de copii jucându-se cu un disc zburător“, i-a pus în dificultate pe toți cercetătorii IA din lume vreme de decenii. Și totuși o echipă de la Google a reușit asta în 2014.⁵ Dacă introduci alt set de pixeli colorați, ți se răspunde „O turmă de elefanți umblând printr-un câmp cu iarbă uscată“ – iarăși corect. Cum a fost posibil? Prin programarea artizanală a algoritmilor, în stil Deep Blue,



Figura 2.11 „Un grup de copii jucându-se cu un disc zburător” – aceasta e descrierea dată de un calculator care nu înțelege nimic despre copii, jocuri sau discuri zburătoare.

pentru a detecta discuri zburătoare, chipuri etc.? Nu, ci prin crearea unei rețele neuronale relativ simple, fără cunoștințe despre lumea fizică sau conținutul ei, și lăsând-o apoi să învețe prin expunerea la cantități imense de date. Jeff Hawkins scria în 2004 că „nici un calculator nu poate [...] vedea la fel de bine ca un șoarece”, dar acele zile sunt de mult apuse.

Așa cum nu înțelegem pe deplin cum învață copiii noștri, încă nu înțelegem pe deplin cum învață aceste rețele neuronale, și de ce uneori dau greș. E clar însă că ele sunt deja extrem de utile și declanșează un val de investiții în învățarea profundă. Învățarea profundă a transformat multe aspecte ale „vederii” calculatorului, de la transcrierea scrisului de mână la analiza video în timp real a automobilelor care se conduc singure. A transformat de asemenea capacitatea calculatoarelor de a transforma vorbirea în text și a o traduce în alte limbi, chiar în timp real – de aceea putem acum vorbi cu asistenți personali digitali cum sunt Siri, Google Now și Cortana. Acele obositoare teste „CAPTCHA”, ca să convingă site-ul că suntem oameni, devin tot mai dificile, pentru a depăși posibilitățile tehnologiei învățării profunde. În 2015, Gogle DeepMind

a lansat un sistem IA folosind învățarea profundă care se putea descurca în zeci de jocuri computerizate la fel cum ar face-o un copil – fără să aibă instrucțiuni, iar curând a învățat să joace mai bine decât orice om. În 2016, aceeași companie a construit AlphaGo, un sistem pentru jocul de go, folosind învățarea profundă ca să evalueze diferite poziții, care l-a învins pe campionul mondial la go. Acest progres stimulează un cerc virtuos, aducând și mai multe finanțări și talente în cercetarea IA, care generează progresul ulterior.

Am examinat în acest capitol natura inteligenței și dezvoltarea ei până acum. Cât va mai dura până ca mașinile să ne depășească în *toate* sarcinile cognitive? Nu știm și trebuie să acceptăm posibilitatea ca răspunsul să fie „niciodată”. Mesajul fundamental al acestui capitol este însă că trebuie să luăm în considerare și posibilitatea că se va întâmpla, poate chiar în cursul vieții noastre. La urma urmei, materia poate fi aranjată așa încât, supunându-se legilor fizicii, să poată memora, calcula și învăța – și nu-i nevoie ca materia să fie biologică. Cercetătorii IA au fost adesea acuzați că promet prea mult și realizează prea puțin, dar adevărul e că unii dintre criticii lor n-au un palmares mai grozav. Unii împing pur și simplu obiectivele mai departe, definind inteligența ca pe ceea ce calculatoarele încă nu pot face, sau ca pe ceea ce ne impresionează. Mașinile sunt acum bune sau excelente la calcule, șah, demonstrarea teoremelor matematice, analiză bursieră, descrierea imaginilor, șofat, jocuri video, go, sinteză vocală, transcrierea vorbirii, traducere și diagnosticarea cancerului, însă unii critici vor strâmba din nas: „Mă rog, dar asta nu-i inteligență adevărată!” Ei ar putea spune că inteligența adevărată presupune doar creșterea munților din peisajul lui Moravec (figura 2.2) care n-au fost încă inundate, la fel cum se spunea înainte că ar trebui să conteze descrierea imaginilor și jocul de go – iar nivelul apei creștea mereu.

Presupunând că nivelul apei va continua să crească, impactul IA asupra societății va fi tot mai mare. Cu mult înainte ca IA să ajungă la nivel uman în toate sarcinile, vom fi confrunțați cu alegeri și provocări fascinante legate de erori, legi, arme și joburi. Care sunt acestea și cum ne putem pregăti cât mai bine pentru a le face față – iată subiectul următorului capitol.

REZUMAT:

- Inteligența, definită drept capacitatea de a îndeplini obiective complexe, nu poate fi măsurată cu un singur IQ, cu un singur spectru de capacități pentru toate obiectivele.
- Inteligența artificială de azi este de regulă *îngustă*, fiecare sistem putând îndeplini numai obiective foarte precise, în vreme ce inteligența umană are un spectru remarcabil de larg.
- Memoria, calculul, învățarea și inteligența sugerează ceva abstract, impalpabil și eteric, fiindcă sunt *independente de substrat*: au o viață proprie care nu depinde de detaliile substratului lor material și nu le reflectă.
- Orice bucată de materie poate fi substratul *memoriei* dacă are mai multe stări stabile.
- Orice materie poate fi *computroniu*, substratul calculului, dacă e compusă din elemente ce pot fi combinate pentru a executa orice funcție. Porțile NAND și neuronii sunt două exemple importante de asemenea „atomi computaționali” universali.
- O rețea neuronală e un substrat puternic pentru *învățare* fiindcă, supunându-se pur și simplu legilor fizicii, se poate rearanja pentru a se auto-perfecționa în efectuarea calculelor dorite.
- Datorită simplității frapante a legilor fizicii, pe noi, oamenii, ne preocupă doar o infimă parte din toate problemele computaționale imaginabile, iar rețelele neuronale sunt de regulă remarcabil de pricepute tocmai la rezolvarea acestei părți minuscule.
- Odată ce tehnologia își dublează puterea, adesea poate fi folosită pentru a proiecta și construi o tehnologie la rândul ei de două ori mai puternică, declanșând dublarea repetată în spiritul legii lui Moore. Timp de peste un secol, prețul tehnologiei informației s-a înjumătățit cam la fiecare doi ani, făcând cu puțință apariția epocii informației.
- Dacă progresul IA continuă, cu mult înainte ca IA să atingă nivelul uman la toate capitolele vom fi confrunțați cu alegeri și provocări fascinante legate de erori, legi, arme și joburi – pe care le vom examina în următorul capitol.

CAPITOLUL 3

Viitorul apropiat: progrese, erori, legi, arme și joburi

*Dacă nu schimbăm repede direcția, ne îndreptăm
spre pieire. — Irwin Corey*

Ce înseamnă să fii om în ziua și în epoca de azi? De pildă, ce prețuim cu adevărat la noi, ce ne deosebește de alte forme de viață și de mașini? Ce prețuiesc la noi alți oameni, așa încât unii vor să ne ofere joburi? Indiferent de răspunsurile noastre la un moment dat, e limpede că ascensiunea tehnologiei le modifică treptat.

Iată, de pildă, cazul meu. Ca om de știință, mă mândresc că îmi stabilesc singur obiectivele, îmi pun la lucru creativitatea și intuiția pentru a aborda o gamă largă de probleme nerezolvate și folosesc limbajul pentru a împărtăși altora ceea ce descopăr. Norocul meu – societatea e dispusă să mă plătească pentru jobul ăsta. În schimb, cu secole în urmă, la fel ca mulți alții, aș fi putut să-mi construiesc personalitatea în jurul meseriei de fermier sau de meșteșugar, dar ascensiunea tehnologiei a redus între timp asemenea profesii la o parte minusculă din forța de muncă. Asta înseamnă că nu oricine își mai poate construi identitatea în jurul muncilor agricole sau meșteșugurilor.

Pe mine unul nu mă deranjează că mașinile de azi mă depășesc în aptitudini manuale precum săpatul și împletitul, fiindcă acestea nu sunt nici hobby-urile mele, nici sursele mele de venit sau de satisfacție personală. Orice iluzii mi-aș fi făcut în privința unor asemenea talente, ele mi-au fost spulberate la vârsta de opt ani,

când școala m-a obligat să urmez lecții de împletit, și n-aș fi luat notă de trecere dacă nu m-ar fi ajutat un binevoitor din clasa a cincea, căruia i s-a făcut milă de mine.

Cum însă tehnologia continuă să se perfecționeze, ascensiunea IA va eclipsa oare în cele din urmă și acele aptitudini care îmi conferă acum stima de sine și valoarea pe piața muncii? Stuart Russell mi-a mărturisit că el și mulți dintre colegii lui cercetători IA trecuseră recent printr-un moment de uluială când au văzut că IA face un lucru la care nu se așteptau. În același spirit, dați-mi voie să vă vorbesc despre câteva dintre momentele mele de uluială, care cred că vestesc depășirea în curând a capacităților umane.

PROGRESE

Agenți de consolidare a învățării profunde

În 2014 am rămas perplex când am văzut un video în care un sistem IA DeepMind învăța să joace jocuri pe calculator. Mai exact, IA juca Breakout (vezi figura 3.1), un joc Atari clasic pe care mi-l aminteam din adolescență. Obiectivul era să manevrezi o paletă așa încât să faci ca o minge să ricoșeze în mod repetat pe un zid de cărămizi; de fiecare dată când loveai o cărămidă, ea dispărea și scorul tău creștea.

Scrisesem și eu pe vremuri câteva jocuri pentru calculator, și știam că nu era greu să scrii un program care să joace Breakout – dar nu asta făcuse echipa DeepMind. Ei creaseră în schimb o IA ca o *tabula rasa*, care nu știa nimic despre acest joc – sau despre vreun altul, sau despre *concepte* cum sunt jocurile, paletele, cărămizile ori mingile. Tot ce primea IA la intervale regulate era o lungă listă de numere: scorul curent și lista numerelor în care noi (dar nu și IA) am recunoaște felul în care erau colorate diferite părți ale ecranului. Tot ce i se spusese era să maximizeze scorul producând la intervale regulate numere în care noi (dar nu și IA) am recunoaște coduri pentru tasta care trebuie apăsată.

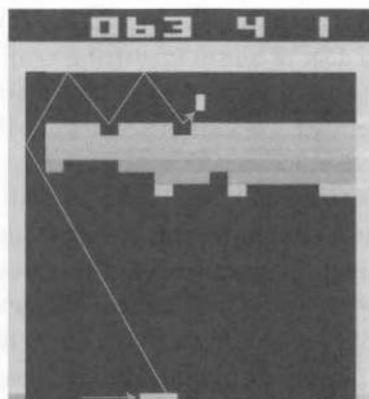


Figura 3.2 După ce a învățat să joace Breakout pornind de la zero, folosind învățarea consolidată pentru a maximiza scorul, IA DeepMind a descoperit strategia optimă: să facă o gaură în extremitatea din stânga a zidului de cărămizi și să lase mingea să tot ricoșeze în spatele lui, adunând rapid puncte. Săgețile reprezintă traiectoriile din trecut ale mingii și paletelor.

La început, IA juca îngrozitor: mișca paleta înainte și înapoi aparent la întâmplare, ratând mai mereu mingea. După un timp, a părut să-și dea seama că mișcarea paletelor spre minge era o idee bună, dar încă rata de cele mai multe ori. Se perfecționa însă pe măsură ce exersa, devenind curând mai bună decât fusesem eu vreodată, returnând fără greș mingea oricât de rapid ar fi venit spre ea. Iar apoi am rămas perplex: a descoperit strategia uimitoare de maximizare a scorului, țintind mereu colțul din stânga-sus pentru a face o gaură în zid și a bloca mingea între partea de sus a zidului și bariera din spatele lui. Era într-adevăr un lucru inteligent. Demis Hassabis avea să-mi spună că programatorii din acea echipă DeepMind nu cunoșteau acest truc, și l-au învățat de la IA pe care o construiseră. Vă recomand să priviți un document video la link-ul pe care vi-l pun la dispoziție.¹

Era aici o trăsătură umană care m-a tulburat: priveam o IA care avea un obiectiv, iar în cele din urmă a învățat să-l îndeplinească mai bine decât creatorii ei. În capitolul precedent am definit inteligența drept capacitatea de a îndeplini obiective complexe, deci

în acest sens IA a lui DeepMind devenea tot mai inteligentă în fața ochilor mei (deși doar în sensul îngust al acestui joc particular). În primul capitol am întâlnit ceea ce informaticienii numesc *agenți inteligenți*: entități care adună informație de la senzori despre mediul lor, iar apoi procesează informația pentru a hotărî cum să reacționeze asupra mediului. Deși IA a lui DeepMind pentru acest joc trăia într-o lume virtuală extrem de simplă, alcătuită din cărămizi, palete și mingi, nu puteam nega că era un agent inteligent.

Cei de la DeepMind și-au publicat curând metoda și și-au dezvăluit codul, explicând că folosea ideea foarte simplă, dar puternică, de învățare consolidată profundă (*deep reinforcement learning*).² Învățarea consolidată elementară e o tehnică clasică de învățare automată, inspirată din psihologia behavioristă, în care primirea unei recompense sporește tendința de a repeta ceva, și viceversa. La fel cum un câine învață să facă giumbușlucuri când asta sporește probabilitatea de a primi imediat încurajări sau hrană de la stăpân, IA a lui DeepMind a învățat să miște paleta pentru a prinde mingea, fiindcă asta sporea probabilitatea de a primi imediat mai multe puncte. DeepMind a combinat această idee cu învățarea profundă: a fost antrenată o rețea neuronală profundă, la fel ca în capitolul precedent, pentru a prezice câte puncte vor fi câștigate în medie apăsând fiecare tastă permisă a claviaturii, iar apoi IA a selectat tasta pe care rețeaua neuronală a găsit-o cea mai promițătoare, dată fiind starea curentă a jocului. În lista trăsăturilor cu care mă mândresc ca om am inclus capacitatea de a aborda o gamă largă de probleme nerezolvate. Să poți juca Break-out și să nu fii în stare de nimic altceva denotă o inteligență foarte îngustă. Pentru mine, ce este cu adevărat important în rezultatul obținut de echipa DeepMind e faptul că învățarea consolidată profundă reprezintă o tehnică absolut generală. Ei au pus exact aceeași IA să exerseze alte 49 de jocuri Atari, iar aceasta a învățat să-și depășească experimentatorii umani în 29 dintre ele, de la Pong la Boxing, Video Pinball și Space Invaders.

N-a durat mult până când aceeași idee de IA să fie testată în jocuri mai moderne, cu lumi tridimensionale, nu doar bidimensionale.

Curând, competitorii din San Francisco ai lui DeepMind în campionatul IA au lansat o platformă numită Universe, unde DeepMind și alți agenți inteligenți pot exersa interacționând cu un întreg computer ca și cum ar fi un joc: apăsând pe orice, tastând orice, deschizând și rulând orice software accesibil – intrând pe un browser și navigând online, de pildă.

Privind spre viitorul învățării consolidate profunde și al perfecționării ei, nu se întrezărește capătul. Potențialul ei nu se mărginește la lumile jocurilor virtuale, căci dacă ai fi robot ai putea privi viața însăși ca pe un joc. Stuart Russell mi-a spus că primul lui mare moment de perplexitate a fost când a văzut robotul Big Dog alergând pe o pantă acoperită cu zăpadă și rezolvând elegant problema locomoției cu picioare la care el muncise ani de-a rândul ca s-o rezolve.³ Iar când în fine obiectivul a fost atins, în 2008, acesta a necesitat o cantitate imensă de muncă a unor programatori iscusiți. După succesul celor de la DeepMind, nu există motive pentru care un robot să nu poată folosi o variantă a învățării consolidate profunde ca să învețe singur să meargă, fără ajutorul unor programatori umani: nu-i nevoie decât de un sistem care să-i dea puncte atunci când face progrese. Roboții din lumea reală ar putea și să învețe să înoate, să zboare, să joace ping-pong, să se lupte și să îndeplinească o listă aproape nesfârșită de alte sarcini motorii, fără ajutor din partea programatorilor umani. Pentru a accelera lucrurile și a reduce riscul de a se bloca sau deteriora în cursul procesului de învățare, ar trebui probabil să parcurgă primele etape ale învățării în realitatea virtuală.

Intuiție, creativitate și strategie

Alt moment definitoriu pentru mine a fost acela când sistemul IA AlphaGo al lui DeepMind a câștigat meciul de go în cinci jocuri împotriva lui Lee Sedol, considerat cel mai bun jucător din lume la începutul secolului XXI.

Se aștepta ca la un moment dat jucătorii umani de go să fie depășiți de mașini, fiindcă aceasta se întâmplase cu colegii lor șahiști în

urmă cu două decenii. Majoritatea specialiștilor în go preziceau însă că va mai trece un deceniu, așa încât triumful lui AlphaGo a fost un moment de răscruce pentru ei, ca și pentru mine. Nick Bostrom și Ray Kurzweil au subliniat cât de dificil era ca IA să progreseze în acest domeniu, ceea ce transpare și din interviurile cu Lee Sedol dinainte și de după ce a pierdut primele trei jocuri:

- octombrie 2015: „Pe baza nivelului constatat [...] cred că voi câștiga clar jocul.”
- februarie 2016: „Am aflat că IA al lui DeepMind de la Google e surprinzător de puternic, și devine și mai puternic, dar sunt încrezător că măcar de data asta pot câștiga.”
- 9 martie 2016: „Sunt foarte surprins, pentru că nu mi-am închipuit că pot pierde.”
- 10 martie 2016: „Am rămas fără cuvinte. [...] Sunt șocat. Recunosc că [...] al treilea joc nu va fi ușor pentru mine.”
- 12 martie 2016: „Mă simt neputincios.”

La un an după ce a jucat cu Lee Sedol, un AlphaGo perfecționat a jucat cu primii douăzeci de jucători din lume, fără să piardă un singur meci.

De ce era pentru mine un lucru atât de important? Mărturisesc că priveam intuiția și creativitatea ca pe două dintre trăsăturile mele intrinsec umane, și, după cum voi arăta imediat, am simțit că AlphaGo le manifesta pe ambele.

Jucătorii de go așază pe rând piese albe și negre pe o tablă de 19×19 (vezi figura 3.2). Există cu mult mai multe poziții de go posibile decât atomii din universul nostru, ceea ce înseamnă că a încerca să analizezi toate șirurile interesante de mutări viitoare devine rapid inutil. Jucătorii se bazează așadar din plin pe intuiția subconștientă pentru a-și completa judecata conștientă, la experți dezvoltându-se o stranie senzație legată de pozițiile puternice și de cele slabe. După cum am văzut în capitoul precedent, rezultatele învățării profunde amintesc uneori de intuiție: o rețea neuronală profundă poate determina că o imagine reprezintă o pisică, fără să poată explica de ce. Echipa DeepMind a pariat deci pe ideea

că învățarea profundă ar putea să recunoască nu doar pisici, dar și poziții puternice de go. Ideea centrală introdusă în AlphaGo a fost de a reuni puterea intuitivă a învățării profunde cu puterea logică a lui GOFAI – acronimul pentru „Good Old-Fashioned AI” („destoinica IA demodată”) dinainte de revoluția învățării profunde. A fost folosită o uriașă bază de date a pozițiilor de go din jocuri dintre oameni, precum și din jocuri în care AlphaGo se confruntase cu o clonă a lui însuși, și a fost antrenată o rețea neuronală profundă pentru a prezice la fiecare poziție probabilitatea ca albul să câștige în cele din urmă. A fost de asemenea antrenată o rețea pentru a prezice următoarele mutări probabile. Apoi, aceste două rețele au fost combinate cu o metodă GOFAI care căuta în mod ingenios într-o listă simplificată de șiruri ulterioare probabile de mutări pentru a identifica următoarea mutare care ar conduce la poziția cea mai puternică.

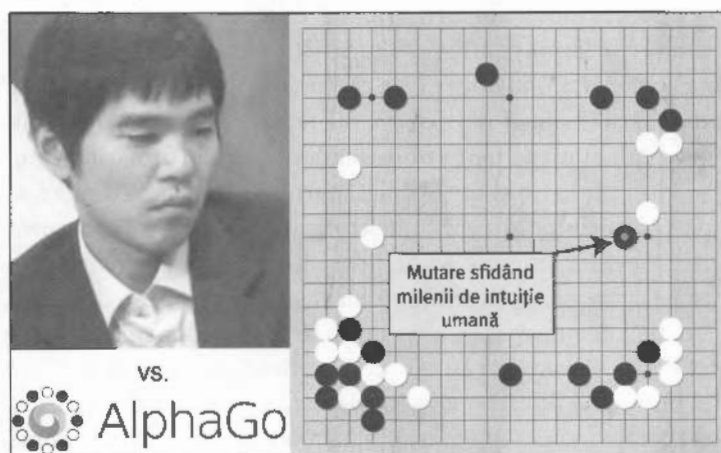


Figura 3.2 IA AlphaGo a lui DeepMind a făcut o mutare extrem de inventivă pe linia 5, sfidând milenii de înțelepciune umană, care peste vreo 50 de mutări s-a dovedit a fi crucială în victoria asupra maestrului go Lee Sedol.

Acest mariaj între intuiție și logică a dat naștere unor mutări care nu erau doar puternice, dar și extrem de inventive. De pildă, milenii de înțelepciune go dictează că la începutul jocului cel mai

bine e să muți pe linia a treia sau a patra de la margine. Dacă muți pe linia a treia, obții pe termen scurt câștig teritorial spre marginea tablei, în schimb, dacă muți pe linia a patra capeti influență strategică pe termen lung spre centru.

La mutarea numărul 37 din al doilea joc, AlphaGo a șocat lumea go-ului, sfidând o veche înțelepciune și jucând pe linia a cincea (figura 3.2), de parcă ar fi fost mai încrezător decât un om în capacitatea de a plănuși pe termen lung, favorizând astfel avantajul strategic în fața celui pe termen scurt. Comentatorii au fost uluiți, iar Lee Sedol chiar s-a sculat și a părăsit încăperea.⁴ După vreo 50 de mutări însă, lupta din colțul stânga-jos al tablei s-a încheiat făcându-se legătura cu piesa neagră de la mutarea 37! Asta a condus la câștigarea jocului, întărind ideea că mutarea lui AlphaGo pe linia a cincea a fost una dintre cele mai inovatoare din istoria go-ului.

Datorită aspectelor lui intuitive și creatoare, go este privit mai curând ca o formă de artă decât ca un joc oarecare. În vechea Chină era considerat una dintre cele patru „arte esențiale”, alături de pictură, caligrafie și muzica *qin*, și a rămas extrem de popular în Asia, unde aproape 300 de milioane de oameni au urmărit primul joc între AlphaGo și Lee Sedol. Ca urmare, lumea go-ului a fost șocată de rezultat și a găsit că victoria lui AlphaGo e un moment crucial pentru omenire. Ke Jie, liderul clasamentului de go la acea vreme, a declarat: „Omenirea a jucat go de mii de ani, dar, după cum ne-a demonstrat IA, abia am ajuns la suprafața lui. [...] Reuniunea jucătorilor umani cu cei computerizați deschide o nouă eră. [...] Împreună, om și IA pot găsi adevărul jocului go.” O asemenea colaborare rodnică om-mașină pare promițătoare în multe domenii, inclusiv în știință, unde putem spera ca IA să ne ajute să înțelegem mai adânc lucrurile și să ne dezvoltăm pe deplin potențialul.

AlphaGo îmi sugerează și o altă idee importantă pentru viitorul apropiat: combinarea intuiției învățării profunde cu logica GOFAI poate produce *strategii* fără egal. Go fiind jocul strategic absolut, se cuvine acum ca IA să concureze (sau să ajute) cele mai bune strategii umane dincolo de tabla de joc – de pildă în strategiile de

investiție, politice și militare. Asemenea probleme strategice din lumea reală sunt de regulă complicate de psihologia umană, lipsa informației și de factori care trebuie modelați aleatoriu, dar sisteme IA pentru jocul de poker au demonstrat deja că nici una dintre aceste provocări nu e insurmontabilă.

Limba naturală

Alt domeniu în care progresele IA m-au uimit recent este limbajul. În tinerețe mi-a plăcut să călătoresc, iar curiozitatea pentru alte culturi și limbi joacă un rol important în identitatea mea. Am vorbit de mic suedeză și engleză, am învățat germană și spaniolă la școală, portugheză și română în urma a două căsnicii, și am învățat singur ceva rusă, franceză și mandarină.

Dar IA a ajuns, după o realizare măreață din 2016, nu există aproape deloc limbi între care eu pot să traduc mai bine decât sistemul IA dezvoltat de echipamentul creierului lui Google.

Ați înțeles ce-am spus? Voiam de fapt să spun următoarele:

Dar IA m-a ajuns din urmă, iar după o realizare importantă din 2016, aproape că n-au rămas limbi între care eu să pot traduce mai bine decât sistemul IA dezvoltat de echipa Google Brain.

Prima dată am tradus fraza în spaniolă, iar pe urmă înapoi în engleză, folosind o aplicație instalată pe laptopul meu acum câțiva ani. În 2016, echipa Google Brain și-a actualizat serviciul gratuit Google Translate pentru a folosi rețele neuronale recurente profunde, iar îmbunătățirea față de mai vechile sisteme GOFAI a fost spectaculoasă:⁶

Dar IA m-a ajuns din urmă, iar după o realizare din 2016, aproape că n-au rămas limbi care să poată traduce mai bine între ele decât sistemul IA dezvoltat de echipa Google Brain.

După cum vedeți, pronumele „eu” s-a pierdut în ocolul prin spaniolă, ceea ce din păcate a schimbat înțelesul. Pe aproape, dar nu tocmai! Trebuie totuși spus, în apărarea IA creată de Google, că adesea sunt criticat fiindcă scriu fraze inutil de lungi, greu de descâlcit, și am ales una dintre cele mai torturate pentru exemplul

meu. Această IA traduce de multe ori impecabil propozițiile tipice. A făcut așadar vâlvă când a apărut, și e suficient de utilă pentru a fi folosită zilnic de milioane de oameni. În plus, grație progreselor în învățarea profundă pentru conversiile vorbire-text și text-vorbire, acești utilizatori pot acum vorbi în telefoanele lor inteligente într-o limbă și asculta rezultatul tradus.

Procesarea limbajului natural este unul dintre domeniile IA cele mai dinamice, iar eu cred că viitoarele succese vor avea un impact puternic, fiindcă limbajul e o trăsătură umană esențială. Cu cât se va perfecționa IA în predicție lingvistică, cu atât se va pricepe mai bine să compună răspunsuri la e-mailuri sau să continue o conversație orală. Asta ar putea da aparența, cel puțin unui novice, că se desfășoară o gândire umană. Sistemele de învățare profundă urmează deci pașii unui prunc pe calea spre faimosul test Turing, în care mașina trebuie să converseze în scris suficient de bine pentru a face pe cineva să creadă că e un om.

IA pentru procesarea limbajului mai are însă o cale lungă de străbătut. Deși mărturisesc că mă simt puțin dezamăgit când sunt tradus de o IA, îmi revin când îmi amintesc că ea nu *înțelege* deloc ceea ce spune. Din antrenamentul cu seturi masive de date, ea descoperă tipare și relații implicând cuvinte, fără să raporteze aceste cuvinte la ceva din lumea reală. De exemplu, ar putea reprezenta fiecare cuvânt printr-o listă cu mii de numere care arată cât de asemănător este cu anumite alte cuvinte. Ar putea conchide de aici că diferența dintre „rege” și „regină” e asemănătoare celei dintre „soț” și „soție” – dar tot n-are vreun indiciu despre ce înseamnă să fii bărbat sau femeie, ori chiar despre faptul că există o realitate fizică, cu spațiu, timp și materie.

Odată ce testul Turing presupune în fond o păcăleală, el a fost criticat pentru că vizează mai curând naivitatea umană decât veritabila inteligență artificială. Dimpotrivă, un test concurent numit *Provocarea Schemei Winograd* țintește drept spre înțelegerea de bun-simț care scapă sistemelor de învățare profundă actuale. Noi, oamenii, apelăm în mod obișnuit la cunoștințe despre lumea reală când analizăm o frază pentru a ne da seama la ce anume se

referă un pronume. De pildă, o provocare Winograd tipică pune întrebarea despre care „ei” este vorba aici:

1. „Consilierii municipali au refuzat să le dea demonstranților aprobarea, fiindcă ei se temeau de violență.”
2. „Consilierii municipali au refuzat să le dea demonstranților aprobarea, fiindcă ei pledau pentru violență.”

Există o competiție anuală IA pentru a răspunde la asemenea întrebări, iar rezultatele sunt deocamdată execrabile.⁷ Această provocare precisă, a înțelege ce anume se referă la ce, l-a spulberat până și pe Google Translate când am înlocuit spaniola cu chineză în exemplul meu de mai sus:

Dar IA m-a ajuns din urmă, după o mare ruptură din 2016, cu aproape nici o limbă, aș putea traduce sistemul IA decât dezvoltat de echipa Google Brain.

Vă invit și pe voi să încercați la <https://translate.google.com>, acum când citiți cartea, și să vedeți dacă IA a lui Google s-a îmbunătățit! Se prea poate să se fi întâmplat asta, căci există abordări promițătoare în mariajul rețelelor neuronale recurente profunde cu GOFAI pentru a crea o IA de procesare a limbajului care include un model al lumii.

Șanse și provocări

Aceste trei exemple au fost doar o mostră, fiindcă IA progresează rapid pe mai multe fronturi importante. Apoi, deși am menționat doar două companii în aceste exemple, grupuri de cercetare din universități și companii rivale nu se află mult în urmă. E mare agitație în catedrele de informatică, odată ce Apple, Baidu, DeepMind, Facebook, Google, Microsoft și alte companii fac oferte tentante studenților, absolvenților și profesorilor.

Exemplele pe care le-am dat n-ar trebui să ne facă să credem că istoria IA constă în perioade de stagnare întrerupte de realizări accidentale. Impresia mea este că asistăm de multă vreme la un progres relativ constant, iar mediile de informare anunță câte o realizare atunci când e trecut pragul care permite o aplicație ce

stârnește imaginația sau un produs util. Cred deci că progresul susținut al IA va continua încă mulți ani. În plus, după cum am văzut în capitolul precedent, nu există vreun motiv fundamental pentru care progresul să nu continue până când IA va egala capacitățile umane la majoritatea capitolelor.

Se pune prin urmare întrebarea: ce efect va avea asta asupra noastră? Cum va modifica progresul IA pe termen scurt condiția umană? Devine tot mai greu să susții că IA este complet lipsită de obiective, anvergură, intuiție, creativitate sau limbaj – trăsături considerate definitorii pentru om. Ca atare, chiar și pe termen scurt, cu mult înainte ca IAG să ne egaleze la toate capitolele, IA ar putea avea un impact puternic asupra felului în care ne privim pe noi înșine, asupra a ceea ce putem face ajutați de IA și asupra profesiunilor care ne rămân disponibile în competiția cu IA. Va fi pozitiv sau negativ acest impact? Ce șanse și provocări pe termen scurt vor apărea?

Tot ce prețuim la civilizație este produsul inteligenței umane, deci dacă o putem spori prin inteligența artificială, ne putem îmbunătăți viața. Chiar și progrese modeste ale IA ar aduce cu sine ameliorări în știință și tehnologie, precum și diminuări corespunzătoare ale accidentelor, maladiilor, in justiției, războaielor, corvezilor și sărăciei. Dar pentru a ne bucura de aceste beneficii ale IA, fără a crea noi probleme, trebuie să răspundem la multe întrebări importante. De exemplu:

1. Cum putem face ca sistemele IA din viitor să fie mai robuste decât cele de azi, așa încât să îndeplinească ce ne dorim fără să cedeze, să funcționeze greșit sau să cadă pradă hackerilor?
2. Cum putem actualiza sistemele noastre juridice pentru a fi mai drepte și mai eficiente, și pentru a ține ritmul cu peisajul digital aflat în schimbare rapidă?
3. Cum putem crea arme mai deștepte și mai puțin dispuse să ucidă civili nevinovați, fără a declanșa o cursă a înarmărilor scăpată de sub control, cu arme autonome letale?
4. Cum ne putem spori prosperitatea prin automatizare, fără a-i lăsa pe oameni lipsiți de venituri și de un rost în viață?

În restul capitolului de față vom cerceta pe rând fiecare dintre aceste întrebări. Ele îi vizează în principal pe informaticieni, juriști, strategii militari și economiști. Pentru a obține însă răspunsurile de care avem nevoie în timp util, este nevoie ca toată lumea să participe la dezbateri, fiindcă, după cum vom vedea, provocările depășesc toate frontierele tradiționale – atât frontierele dintre profesii, cât și cele dintre națiuni.

ERORI VS IA ROBUSTĂ

Tehnologia informației a avut deja un mare impact pozitiv practic asupra fiecărei activități umane – de la știință la finanțe, producție, transport, ocrotirea sănătății, energie și comunicații –, iar acest impact pălește în comparație cu progresul pe care l-ar putea aduce IA. Dar cu cât ajungem să ne bazăm mai mult pe tehnologie, cu atât mai important devine ca ea să fie robustă și demnă de încredere, făcând ceea ce vrem noi să facă.

În decursul istoriei umane, am învățat din greșeli pentru a ne menține benefică tehnologia. Am inventat focul, care ne-a dat adesea bătăi de cap, apoi am inventat extingtorul, ieșirea de incendiu, alarma de incendiu și serviciul de pompieri. Am inventat automobilul, care a produs adesea accidente, apoi am inventat centurile de siguranță, air bag-urile și automobilele care se conduc singure. Până acum, tehnologiile noastre au provocat în genere accidente suficient de puține și de limitate pentru ca neajunsurile să fie eclipsate de avantaje. Cum însă elaborăm inevitabil tehnologii și mai puternice, vom ajunge în punctul în care chiar și un singur accident poate fi atât de catastrofal încât să eclipseze toate avantajele. Unii dau drept exemplu războiul nuclear global accidental, alții pandemia produsă de bioinginerie, iar în capitolul următor vom vorbi despre temerea ca IA să conducă la dispariția omenirii. Nu-i nevoie totuși de asemenea exemple extreme pentru a ajunge la o concluzie esențială: pe măsură ce tehnologia devine tot mai puternică, trebuie să ne bazăm tot mai puțin pe strategia

de a învăța din greșeli. Cu alte cuvinte, *trebuie să fim mai curând proactivi decât reactivi*, investind în cercetările legate de siguranță pentru a evita ca accidente să se producă fie și o singură dată. De aceea societatea investește mai mult în siguranța reactoarelor nucleare decât în siguranța curselor de șoareci.

Acesta e și motivul pentru care, după cum am văzut în capitolul 1, la conferința din Puerto Rico s-a manifestat un puternic interes colectiv pentru cercetările în siguranța IA. Computerele și sistemele IA au cedat mereu, dar de data asta lucrurile stau altfel: IA pătrunde treptat în lumea reală, și nu e o simplă neplăcere dacă provoacă prăbușirea rețelei electrice, a bursei sau a sistemului armamentului nuclear. În restul acestei secțiuni voi vorbi despre cele patru mari domenii ale cercetărilor tehnice în siguranța IA care domină dezbaterile actuale și sunt studiate pretutindeni în lume: *verificare, validare, securitate și control*.^{*} Pentru a evita să devenim plicticoși și arizi, vă propun să examinăm succesele și eșecurile tehnologiei informației din trecut în diferite sectoare, precum și învățămintele pe care le putem trage din ele și provocările pe care le prezintă.

Deși majoritatea acestor povești sunt vechi, implicând sisteme computerizate de mult depășite, pe care aproape nimeni nu le-ar caracteriza drept IA, și au provocat pagube neînsemnate, vom vedea că putem trage din ele învățăminte prețioase pentru proiectarea unor viitoare sisteme IA sigure și puternice, ale căror eșecuri s-ar putea dovedi cu adevărat catastrofale.

IA pentru explorarea spațiului

Să începem cu un subiect care mi-e drag: explorarea spațiului. Tehnologia calculatorului ne-a permis să trimitem oameni pe Lună și nave fără echipaj uman să exploreze toate planetele sistemului solar, ba chiar și să coboare pe suprafața lui Titan, satelitul

^{*} Dacă vreți o hartă mai amănunțită a peisajului cercetărilor în siguranța IA, există una interactivă, elaborată printr-un efort comun dirijat de Richard Mallah de la FLI: <https://futureoflife.org/landscape>. (N. a.)

lui Saturn, și pe o cometă. După cum vom vedea în capitolul 6, în viitor IA ar putea ajuta la explorarea altor sisteme solare și galaxii – cu condiția să fie lipsită de erori. Pe 4 iunie 1996, oamenii de știință care sperau să cerceteze magnetosfera Pământului sărbătoreau lansarea rachetei Ariane 5 a Agenției Spațiale Europene, având la bord instrumentele construite de ei. Peste 37 de secunde însă le-a pierit zâmbetul, căci racheta a explodat într-un foc de artificii care costase sute de milioane de dolari.⁸ Cauza era un software defect, care opera cu un număr prea mare pentru a intra în cei 16 biți alocați lui.⁹ Doi ani mai târziu, sonda spațială Mars Climate Orbiter, lansată de NASA, a intrat din greșeală în atmosfera Planetei Roșii și s-a dezintegrat fiindcă două părți ale software-ului foloseau unități diferite pentru forță, provocând o eroare de 445% în sistemul de control al motorului.¹⁰ Era a doua eroare scump plătită de NASA: misiunea Mariner 1 către planeta Venus a explodat după lansarea de la Cape Canaveral pe 22 iulie 1962, după ce software-ul pentru controlul zborului a fost dat peste cap de un semn de punctuație incorect.¹¹ Parcă pentru a dovedi că nu numai occidentalii stăpâneau arta de a lansa erori în spațiu, misiunea sovietică Phobos 1 a eșuat pe 2 septembrie 1988. Era cea mai grea navă interplanetară lansată vreodată, cu obiectivul ambițios de a depune un modul pe Phobos, satelitul lui Marte – dar absența unei cratime a provocat transmiterea comenzii „sfârșit-de-misiune” către nava aflată în drum spre Marte, oprindu-i toate sistemele.¹²

Ce aflăm din aceste exemple este cât de important e ceea ce informaticienii numesc *verificare*: să te asiguri că software-ul satisface pe deplin toate cerințele. Cu cât sunt în joc mai multe vieți și resurse, cu atât avem nevoie de o încredere mai mare că software-ul va funcționa așa cum dorim. Din fericire, IA poate ajuta la automatizare și poate îmbunătăți procesul de verificare. De exemplu, un nucleu al sistemului de operare universal numit *seL4* a fost recent verificat matematic, oferind o garanție puternică împotriva defectării și a operațiilor periculoase: deși nu a fost anunțat cu surle și trâmbițe ca Microsoft Windows și Mac OS,

poți fi sigur că nu vei primi mesaje de eroare precum „ecranul albastru al morții“ („*the blue screen of death*“). Agenția pentru Proiecte de Cercetare Avansată pentru Apărare (Defence Advanced Research Projects Agency – DARPA) din Statele Unite a sponsorizat dezvoltarea unui set de instrumente de siguranță cu sursă deschisă numit HACMS (*high-assurance cyber military systems* – sisteme militare cibernetice de mare siguranță) care s-a dovedit a fi sigur. O provocare importantă este de a face ca aceste instrumente să devină suficient de puternice și ușor de folosit, așa încât să fie larg răspândite. O altă provocare ține de faptul că însăși sarcina verificării devine mai dificilă când software-ul pătrunde în roboți și în noi medii, iar software-ul preprogramat tradițional este înlocuit de sisteme IA care învață continuu, deci își schimbă comportamentul, după cum am văzut în capitolul 2.

IA pentru finanțe

Finanțele sunt alt domeniu care a fost transformat de tehnologia informației, permițând realocarea eficientă a resurselor oriunde în lume cu viteza luminii, și înlesnind finanțarea avantajoasă pentru orice, de la ipotecă la companii startup. Progresul IA e probabil să ofere în viitor mari posibilități de profit din tranzacții financiare: majoritatea deciziilor de cumpărare/vindere a acțiunilor sunt acum făcute automat de calculatoare, iar absolvenții mei de la MIT sunt tentați cu salarii astronomice pentru a îmbunătăți tranzacționarea algoritmică.

Verificarea e importantă și pentru software-ul financiar, ceea ce a aflat pe pielea ei firma americană Knight Capital pe 1 august 2012, pierzând 440 de milioane de dolari în 45 de minute, după ce a instalat un software pentru tranzacții neverificate.¹³ „Crahul fulger“ de un bilion de dolari din 6 mai 2010 merită menționat din alt motiv. Deși a provocat haos vreme de o jumătate de oră, până când piețele s-au stabilizat, prețul acțiunilor unor companii mari precum Procter & Gamble fluctuând între un penny și 100 000 de dolari,¹⁴ problema n-a fost declanșată de erori sau disfuncții ale

calculatoarelor pe care verificarea le-ar fi putut evita. Ea a avut în schimb drept cauză încălcarea așteptărilor: programele de tranzacționare automată de la numeroase companii au ajuns să opereze într-o situație neprevăzută, în care presupunerile lor nu mai erau valabile – de pildă, presupunerea că, dacă un computer al unei burse anunța că o acțiune are prețul de un cent, atunci acțiunea valora într-adevăr un cent.

Crahul fulger ilustrează importanța a ceea ce informaticienii numesc *validare*: dacă verificarea pune întrebarea „Am construit oare corect sistemul?“, validarea pune întrebarea „Am construit oare sistemul corect?“* De exemplu: nu cumva sistemul se bazează pe presupuneri care s-ar putea să nu fie întotdeauna valabile? Dacă da, atunci cum poate fi perfecționat pentru a trata mai bine incertitudinea?

IA pentru industrie

Evident, IA deține un potențial imens pentru îmbunătățirea producției, controlând roboți care sporesc deopotrivă eficiența și precizia. Imprimantele 3-D pot face acum prototipuri pentru orice, de la blocuri la dispozitive micromecanice mai mici decât un grăunte de sare.¹⁵ În vreme ce roboți industriali uriași construiesc automobile și avioane, freze și strunguri controlate de calculator, care pot fi cumpărate la prețuri accesibile, utilează nu doar fabricile, ci și cele peste o mie de mici ateliere-laborator ale comunităților locale din lumea întreagă, unde entuziaștii își pot materializa ideile.¹⁶ Cu cât avem însă mai mulți roboți în jurul nostru, cu atât e mai important să verificăm și să validăm software-ul lor. Primul om ucis de un robot a fost Robert Williams, muncitor la uzinele Ford din Flat Rock, Michigan. În 1979, un robot care trebuia să coboare piese dintr-un spațiu de depozitare s-a blocat, iar Williams s-a urcat să ia singur piesele. Robotul a început să funcționeze pe tăcute și i-a zdrobit capul, continuând să lucreze o jumătate de oră,

* Mai exact, verificarea întreabă dacă sistemul îndeplinește specificațiile, iar validarea dacă au fost alese specificațiile corecte. (N. a.)

până când ceilalți muncitori au descoperit ce se întâmplase.¹⁷ Urmatorea victimă a roboților a fost Kenji Urada, inginer de întreținere la uzinele Kawasaki din Akashi, Japonia. În 1981, în timp ce lucra la un robot defect, i-a atins din greșeală comutatorul și a fost zdrobit de brațul hidraulic al robotului.¹⁸ În 2015, un muncitor de la uzinele Volkswagen din Baunatal, Germania, pune în funcțiune un robot care trebuia să apuce piesele auto și să le manevreze. Ceva s-a defectat, robotul l-a apucat și l-a strivit de un banc metalic.¹⁹

Deși aceste accidente sunt tragice, e important de observat că ele reprezintă o fracțiune minusculă din ansamblul accidentelor industriale. De fapt, accidentele industriale nu s-au înmulțit, ci s-au *rărit* odată cu perfecționarea tehnologiei, scăzând în Statele Unite de la circa 14 000 de decese în 1970 la 4 821 în 2014.²⁰ Cele trei accidente pomenite aici arată că dotarea unor mașini comune cu inteligență ar trebui să sporească siguranța industrială, învățându-i pe roboți să fie mai atenți în apropierea oamenilor. Toate cele trei accidente puteau fi evitate printr-o validare mai bună: roboții au produs vătămări nu din rea voință sau din pricina vreunei erori din program, ci fiindcă au făcut presupuneri greșite – că persoana nu se afla acolo sau că era o piesă auto.



Figura 3.3 Dacă roboții industriali sunt scumpi și dificil de programat, există acum o tendință către roboți mai ieftini, echipați cu IA, care pot învăța ce să facă de la muncitori fără experiență în programare.

IA pentru transport

Dacă IA poate salva multe vieți în industrie, ea poate salva încă și mai multe vieți în transport. Numai accidentele de mașină au ucis peste 1,2 milioane de oameni în 2015, iar accidentele aviatice, feroviare și navale au ucis împreună alte câteva mii. În Statele Unite, unde sunt standarde înalte de siguranță, accidentele vehiculelor motorizate au făcut în 2016 circa 35 000 de victime, de șapte ori mai mult decât toate accidentele industriale la un loc.²¹ Într-o discuție pe această temă la întrunirea anuală a Asociației pentru Progresul Inteligenței Artificiale din 2016, desfășurată la Austin, Texas, informaticianul israelian Moshe Vardi a spus că IA nu doar *poate* reduce accidentele auto, dar *trebuie* s-o facă: „Este un imperativ moral!” a exclamat el. Cum majoritatea accidentelor de mașină sunt provocate de erori umane, se crede că automobilele echipate cu IA care se conduc singure pot elimina cel puțin 90% din decesele rutiere, iar acest optimism încurajează apariția efectivă a unor asemenea mașini pe drumuri. Elon Musk estimează că viitoarele automobile care se conduc singure vor fi nu doar mai sigure, dar vor aduce și bani proprietarilor lor atunci când aceștia n-au nevoie de ele, putând concura cu Uber și Lyft.

Până acum, automobilele care se conduc singure s-au dovedit mai sigure decât șoferii umani, iar accidentele petrecute subliniază importanța și dificultatea validării. Prima ciocnire minoră provocată de un asemenea automobil Google a avut loc pe 14 februarie 2014, din cauza unei presupuneri incorecte privind un autobuz: că șoferul va ceda atunci când mașina a apărut în fața lui. Primul accident mortal provocat de un automobil Tesla, care a intrat în remorca unui camion pe 7 mai 2016, a fost provocat de două presupuneri greșite:²² că latura albă strălucitoare a remorcii era o parte a cerului strălucitor și că șoferul (care se pare că urmărea un film cu Harry Potter) era atent și ar fi intervenit în caz de urgență.*

Uneori însă, buna verificare și validarea nu sunt de ajuns pentru a evita accidentele, fiindcă avem nevoie și de un bun *control*:

* Chiar luând în calcul acest accident, introducerea autopilotului Tesla a redus cu 40% numărul accidentelor: <http://tinyurl.com/teslasafety>.

capacitatea operatorului uman de a urmări sistemul și de a-i schimba comportamentul dacă e necesar. Pentru ca asemenea sisteme cu „om în buclă” să funcționeze bine e esențială eficacitatea comunicării om-mașină. În această idee, o lumină roșie pe tabloul de bord te va alerta dacă din greșeală ai lăsat deschis portbagajul mașinii. Când feribotul britanic pentru automobile Herald of Free Enterprise a părăsit portul Zeebrugge, pe 6 martie 1987, cu rampa de la provă coborâtă, n-a existat vreo lumină de alarmă sau alt semnal de avertizare pentru căpitan, iar feribotul s-a răsturnat în scurt timp, făcând 193 de victime.²³

Altă greșeală tragică de control care putea fi evitată printr-o mai bună comunicare om-mașină s-a petrecut în noaptea de 1 iunie 2009, când zborul Air France 447 s-a prăbușit în Oceanul Atlantic, toți cei 228 de oameni aflați la bord murind. Conform raportului oficial, „echipajul n-a înțeles că pierdeau din portanță, și deci n-a aplicat manevra de redresare” – coborârea botului avionului – decât prea târziu. Experții în securitatea aviatică au afirmat că accidentul putea fi evitat dacă în cabină ar fi existat un indicator al „unghiului de atac” care să le arate piloților că botul era îndreptat prea mult în sus.²⁴

Când zborul Air France 148 s-a prăbușit în Munții Vosgi, în apropiere de Strasbourg, pe 20 ianuarie 1992, 87 de persoane pierzându-și viața, cauza n-a fost lipsa de comunicare om-mașină, ci o interfață derutantă. Piloții au introdus „33” pe tastatură, pentru că voiau să coboare la un unghi de 3,3 grade, dar pilotul automat a interpretat că asta reprezintă 3 300 de picioare pe minut, fiindcă se afla într-un mod diferit, iar ecranul era prea mic pentru a afișa modul și a permite piloților să-și dea seama de greșeală.

IA pentru energie

Tehnologia informației a făcut minuni pentru generarea și distribuția energiei, cu algoritmi sofisticati care echilibrează producția și consumul în rețelele electrice, și cu sisteme de control sofisticate care mențin funcționarea sigură și eficientă a centralelor

electrice. Viitoarele progrese ale IA vor face probabil și mai inteligentă „rețeaua inteligentă“, pentru a se adapta optim cererii și ofertei schimbătoare până la nivelul individual al panourilor solare de pe acoperișuri și al sistemelor casnice de stocare a energiei. Totuși, pe 14 august 2003, o pană de curent a afectat circa 55 de milioane de americani și canadieni, mulți dintre ei rămânând fără energie câteva zile. Și aici, cauza principală a fost comunicarea defectuoasă om-mașină: o eroare în software a împiedicat sistemul de alarmă dintr-o cameră de comandă din Ohio să alerteze operatorii că e nevoie să redistribuie energia înainte ca o problemă minoră (linii de transmisie supraîncărcate prinse într-un frunziș des) să ia amploare și să scape de sub control.²⁵

Topirea parțială a reactorului nuclear din centrala Three Mile Island, Pennsylvania, din 28 martie 1979, a adus pierderi de un miliard de dolari și a declanșat o reacție violentă împotriva energiei nucleare. Raportul final asupra accidentului a identificat mai multe cauze, între care confuzia provocată de slaba comunicare om-mașină.²⁶ În particular, becul de avertizare despre care operatorii credeau că arată dacă o valvă de siguranță e închisă sau deschisă, indica doar dacă a fost transmis un semnal pentru a închide valva – așa încât operatorii nu și-au dat seama că valva rămăsese înțepenită deschisă.

Aceste accidente legate de energie și transmiterea ei ne arată că, pe măsură ce punem IA să controleze și mai multe sisteme fizice, trebuie să facem eforturi de cercetare serioase vizând nu doar funcționarea corectă a mașinilor pe cont propriu, dar și colaborarea eficientă cu controlorii lor umani. Progresele IA vor aduce cu sine atât crearea unor interfețe perfecționate pentru utilizator în vederea partajării informației, cât și soluții pentru alocarea optimă a sarcinilor în echipele om-calculator – de exemplu, identificarea situațiilor în care controlul trebuie transferat, pentru aplicarea eficientă a judecății umane în deciziile cruciale, în loc să distragi controlorii umani cu un flux de informații neimportante.

IA pentru ocrotirea sănătății

IA are un potențial imens pentru îmbunătățirea asistenței medicale. Digitalizarea fișelor medicale a permis deja doctorilor și pacienților să ia hotărâri mai bune și mai rapide, și să primească instantaneu ajutorul experților de oriunde în lume la diagnosticarea imaginilor digitale. Cei mai buni experți în acest tip de diagnostic ar putea deveni curând sistemele IA, dat fiind progresul rapid în imagistica computerizată și învățarea profundă. De pildă, un studiu olandez din 2015 arăta că diagnosticarea computerizată a cancerului de prostată folosind imagistica prin rezonanță magnetică (IRM) era la fel de bună ca cel al radiologilor umani,²⁷ iar un studiu Stanford din 2016 arăta că IA poate diagnostica mai bine decât anatomopatologii umani cancerul pulmonar, folosind imaginile de la microscop.²⁸ Dacă învățarea automată ar putea ajuta la descoperirea relațiilor dintre gene, boli și răspunsul la tratament, ar revoluționa actul medical, ar îmbunătăți sănătatea animalelor din ferme și ar duce la obținerea de recolte mai rezistente. Apoi, roboții ar putea deveni chirurghi mai preciși și mai de încredere decât oamenii, chiar și fără a folosi IA avansată. O mare diversitate de intervenții chirurgicale robotice au fost efectuate cu succes în ultimii ani, precizia, miniaturizarea și inciziile mai mici făcând să scadă pierderile de sânge, durerile și perioada de recuperare.

Din nefericire însă, câteva cazuri tragice ne atrag atenția asupra importanței unui software robust și în domeniul sănătății. De exemplu, instalația canadiană de radioterapie Therac-25 a fost proiectată pentru a trata pacienții cu cancer în două moduri: fie cu un fascicul de electroni de putere scăzută, fie cu un fascicul de raze X de mare putere, un megavolt, concentrat doar asupra țintei cu ajutorul unui ecran protector special. O eroare dintr-un software neverificat i-a făcut pe tehnicieni să trimită fasciculul de un megavolt, când ei credeau că administrează fasciculul de putere scăzută, și fără ecranul protector, provocând decesul câtorva pacienți.²⁹ Încă și mai mulți pacienți au murit din cauza supradozei de radiații la Institutul Oncologic Național din Panama, unde

echipamentul de radioterapie cu cobalt-60 a fost programat cu perioade de expunere prea lungi în 2000 și 2001, din cauza unei interfețe care nu fusese corect validată.³⁰ Conform unui raport recent,³¹ în Statele Unite, între 2000 și 2013, accidentele provocate de chirurgia robotizată au dus la 144 de decese și 1 391 de vătămări, problemele fiind legate nu numai de hardware – arcuri electrice și arderi, sau piese desprinse din instrumente căzând în pacient –, dar și de software – mișcări necontrolate și deconectări spontane.

Vestea bună este că restul celor aproape două milioane de intervenții chirurgicale robotizate incluse în raport s-au desfășurat perfect, iar roboții par să fie destul de siguri în chirurgie. Conform unui studiu guvernamental american, malpraxisul medical în spitale are drept efect peste 100 000 de decese anual numai în Statele Unite,³² prin urmare imperativul moral de a dezvolta IA pentru medicină e probabil încă mai puternic decât pentru automobile care se conduc singure.

IA pentru comunicații

Industria comunicațiilor este pesemne cea în care calculatoarele au avut până acum cel mai mare impact. După introducerea tablourilor de comandă telefonice computerizate în anii '50, a internetului în anii '60 și a World Wide Web în 1989, miliarde de oameni comunică, fac cumpărături, citesc știri, se uită la filme sau joacă jocuri online, obișnuți să aibă informația din întreaga lume la un clic distanță – și adesea gratuit. Acel *internet al lucrurilor* (*internet of things*) aflat în plin avânt promite să crească eficiența, precizia, comoditatea și beneficiul economic aducând online totul, de la lămpi, termostate și frigidere la transpondere cu biocipuri și animale de fermă.

Aceste succese spectaculoase în conectarea lumii i-au adus pe informaticieni în fața unei a patra provocări: trebuie să îmbunătățească nu numai verificarea, validarea și controlul, dar și *securitatea* împotriva software-ului malefic (malware) și a hackerilor. Dacă problemele menționate mai sus erau rezultatul unor greșeli

neintenționate, securitatea vizează acțiuni *deliberate*. Primul malware care a atras atenția mediilor de informare a fost așa-numitul „vierme informatic Morris“, lansat pe 2 noiembrie 1988, care exploata erorile din sistemul de operare UNIX. A fost, chipurile, o încercare greșit orientată de a număra câte calculatoare erau online, și deși a infectat și blocat circa 10% din cele 60 000 de calculatoare care compuneau pe atunci internetul, asta nu l-a împiedicat pe creatorul său, Robert Morris, să ajungă în cele din urmă profesor de informatică la MIT.

Alte malware-uri exploatează vulnerabilitățile nu doar ale software-ului, ci și ale oamenilor. Pe 5 mai 2000, parcă pentru a sărbători ziua mea de naștere, lumea a primit e-mailuri cu subiectul „ILOVEYOU“ (TEIUBESC) de la cunoștințe și colegi, iar acei utilizatori ai Microsoft Windows care au deschis documentul atașat „LOVE-LETTER-FOR-YOU.txt.vbs“ (SCRISOARE-DE-DRAGOSTE-PENTRU-TINE) au lansat fără să vrea un program care le-a avariat calculatorul și a retrimis e-mailul tuturor din lista lor de adrese. Creat de doi tineri programatori din Filipine, acest vierme informatic a infectat aproximativ 10% din internet, la fel cum o făcuse viermele lui Morris, dar fiindcă internetul era deja mult mai vast, a devenit una dintre cele mai grave infectări din toate timpurile, afectând peste 50 de milioane de calculatoare și provocând o pagubă de peste 5 miliarde de dolari. După cum din nefericire știți, internetul e în continuare infestat cu nenumărate tipuri de malware, pe care experții în securitate îl clasifică în viermi, troieni, viruși și alte categorii cu rezonanțe înspăimântătoare, iar daunele provocate se întind pe un spectru larg, de la afișarea unor mesaje inocente la ștergerea fișierelor, furtul informațiilor personale, spionare și deturnarea calculatorului pentru a trimite spamuri.

Dacă malware-ul vizează orice calculator găsește, *hackerii* atacă anumite ținte de interes – între care, recent, Target, TJ Maxx, Sony Pictures, Ashley Madison, compania petrolieră saudită Aramco și Comitetul Național Democrat al Statelor Unite. Jaful însă pare să fie și mai spectaculos. În 2008, *hackerii* au furat 130 de milioane de numere ale cărților de credit și alte informații despre cont de

la sistemele de plată Heartland, iar în 2013 au pătruns în peste un miliard (!) de conturi de e-mail Yahoo!.³³ În 2014, de la Oficiul Guvernamental de Management al Personalului din Statele Unite s-au sustras informații din dosarele personale și cererile de angajare a peste 21 de milioane de oameni, între care angajați cu acces la informații confidențiale, și au fost obținute amprentele unor agenți sub acoperire.

Mă crucesc de fiecare dată când citesc despre un sistem prezentat a fi 100% sigur și inatacabil. Totuși, viitoarele sisteme IA trebuie să fie „inatacabile” înainte să preia comanda infrastructurii critice sau a armamentului, de pildă – așa încât creșterea rolului IA în societate ridică miza securității calculatoarelor. Dacă unii hackeri exploatează naivitatea umană sau vulnerabilitățile complexe din software-ul recent lansat, alții se conectează neautorizat la calculatoare străine speculând erori simple de programare care au rămas mult timp neobservate. Bug-ul „Heartbleed” a supraviețuit din 2012 până în 2014 într-una dintre cele mai populare biblioteci de software pentru comunicații sigure între calculatoare, iar bug-ul „Bashdoor” s-a aflat chiar în sistemul de operare al calculatoarelor Unix din 1989 până în 2014. Asta înseamnă că instrumentele IA pentru ameliorarea verificării și validării vor spori și securitatea.

Din păcate, sisteme IA perfecționate pot fi folosite și pentru a găsi noi vulnerabilități și a-i înzestra pe hackeri cu arme și mai sofisticate. Închipuie-ți, de exemplu, că primești un e-mail-capcană neobișnuit de personalizat care încearcă să te convingă să divulgi informații personale. E trimis de pe contul prietenei tale de către IA a unui hacker și îi imită stilul pe baza analizei altor mesaje trimise de ea, incluzând o mulțime de informații personale despre tine din alte surse. Te-ai lăsa păcălit? Dar dacă mesajul-capcană ar părea să provină de la compania care ți-a eliberat cardul de credit și e urmat de un apel telefonic în care îți vorbește o voce umană prietenoasă despre care nu știi că e generată prin IA? În cursa înarmărilor dintre atac și apărare din securitatea informatică nu prea avem indicii că apărarea ar câștiga.

Noi, oamenii, suntem animale sociale care am subjugat celelalte specii și am cucerit Pământul grație capacității noastre de a coopera. Am conceput legi pentru a stimula și înlesni cooperarea, așa încât, dacă IA poate îmbunătăți sistemele noastre juridice și de guvernare, ne-ar ajuta să cooperăm încă și mai eficient decât până acum și să ne atingem potențialul maxim. Și există o mulțime de lucruri de îmbunătățit, atât în aplicarea, cât și în elaborarea legilor, așa încât vă propun să le examinăm pe ambele.

Ce îți vine în minte când te gândești la sistemul juridic din țara ta? Termene lungi, costuri ridicate și, uneori, nedreptate – și nu ești singurul. N-ar fi oare minunat dacă primele cuvinte care ți-ar veni în minte ar fi „eficiență” și „corectitudine”? Din moment ce procesul juridic poate fi privit în mod abstract ca un calcul, introducând informații despre dovezi și legi și extrăgând o decizie, unii visează să-l automatizeze complet cu *robojudecători*: sisteme IA care aplică neobosit aceleași standarde juridice înalte la orice hotărâre judecătorească, fără a cădea pradă erorilor umane cum sunt părtinirea, oboseala și ignorarea celor mai recente cunoștințe tehnice.

Robojudecători

Byron De La Beckwith jr. a fost condamnat în 1994 pentru asasinarea, în 1963, a militantului pentru drepturile civile Medgar Evers, dar două jurii diferite din Mississippi, compuse numai din albi, îl achitaseră la un an după crimă, deși dovezile fizice erau în esență aceleași.³⁴ Din nefericire, istoria justiției e plină de asemenea verdicte influențate de culoarea pielii, gen, orientare sexuală, religie, naționalitate și alți factori. În principiu, robojudecătorii ar putea face ca pentru prima dată în istorie toți oamenii să fie cu adevărat egali în fața legii: ei ar fi programați să fie identici și să trateze pe toată lumea în mod egal, aplicând transparent legea, fără părtinire.

Robojudecătorii ar putea elimina și preferințele umane accidentale, neintenționate. De pildă un controversat studiu asupra judecătorilor israelieni afirma că aceștia dau verdicte semnificativ mai aspre atunci când le este foame: dacă imediat după micul dejun refuză 35% din cererile de eliberare condiționată, chiar înainte de prânz ei refuză 85%.³⁵ Alt neajuns al judecătorilor umani este că s-ar putea să nu aibă timp să studieze toate detaliile unui proces. Robojudecătorii pot fi însă cu ușurință copiați, fiindcă în esență ei nu sunt decât un software, așa încât toate procesele aflate în așteptare pot fi judecate în paralel, fiecare având propriul lui robojudecător pentru atât timp cât e nevoie. În fine, dacă judecătorilor umani le e imposibil să stăpânească toate cunoștințele tehnice cerute de fiecare caz în parte, de la dispute spinoase pentru brevete la misterele unor crime depinzând de cele mai avansate cunoștințe de medicină legală, viitorii robojudecători vor avea memorie și capacitate de învățare practic nelimitate.

Într-o bună zi, asemenea robojudecători ar putea fi deci mai eficienți și mai corecți grație nepărtinirii, competenței și transparenței. Eficiența i-ar face să devină încă și mai corecți: accelerând procesul și îngreunând încercările avocaților versați de a influența verdictul, actul de justiție ar deveni mult mai ieftin, ceea ce ar spori șansele ca un om sărac sau o companie startup să câștige în fața unui miliardar sau a unei corporații multinaționale pentru care lucrează o armată de avocați.

Dar ce se întâmplă dacă robojudecătorii au bug-uri sau sunt victimele hackerilor? Ambele primejdii au afectat deja mașinile de vot automat, iar atunci când miza e închisoarea sau o sumă enormă, riscurile unui atac cibernetic sunt și mai mari. Chiar dacă IA devine suficient de robustă pentru a ne da încredere că robojudecătorul folosește algoritmul legal, va avea oare toată lumea senzația că înțelege îndeajuns logica raționamentelor lui pentru a-i respecta verdictul? Această problemă e accentuată de recentul succes al rețelelor neuronale, care adesea depășesc cu mult algoritmi IA tradiționali ușor de înțeles, prețul plătit fiind neinteligibilitatea. Dacă inculpații vor să știe *de ce* au fost condamnați, nu

ar avea oare ei dreptul la un răspuns mai bun decât „am antrenat sistemul cu o mulțime de date, iar asta a hotărât”? În plus, studii recente au arătat că dacă antrenezi un sistem neuronal de învățare profundă cu cantități mari de date privind deținuții, el poate prezice cine e probabil să recidiveze (și ar trebui deci să i se refuze eliberarea condiționată) mai bine decât judecătorii umani. Ce se întâmplă însă dacă sistemul găsește că recidivismul e corelat statistic cu sexul sau rasa deținutului – ar fi considerat un robojudecător sexist sau rasist care trebuie reprogramat? Un studiu din 2016 susține că software-ul folosit în Statele Unite pentru predicția recidivismului îi defavoriza pe afro-americani și a contribuit la sentințe nedrepte.³⁶ Acestea sunt aspecte importante care trebuie analizate și discutate pentru a ne asigura că IA rămâne benefică. Nu suntem puși în fața unei hotărâri de tipul totul-sau-nimic în privința robojudecătorilor, ci ar trebui să hotărâm în ce măsură și în ce ritm vrem să folosim IA în sistemul nostru juridic. Vrem ca judecătorii umani să fie ajutați de sisteme de decizie bazate pe IA, la fel ca medicii de mâine? Vrem să mergem mai departe și să avem decizii ale robojudecătorilor, iar apelul să fie lăsat în seama judecătorilor umani, sau vrem să mergem până la capăt, dând mașinilor ultimul cuvânt chiar și în cazul pedepsei cu moartea?

Controverse juridice

Am examinat până acum doar *aplicarea legii*, să ne întoarcem la *conținutul ei*. Există un larg consens cu privire la faptul că legile noastre trebuie să evolueze pentru a ține ritmul cu tehnologia. De exemplu, cei doi programatori care au creat viermele ILOVEYOU, provocând pagube de miliarde de dolari, au fost achitați de toate acuzațiile pentru că în Filipine nu existau pe atunci legi în privința malware-ului. De când ritmul progresului tehnologic s-a accelerat, legile trebuie actualizate și mai rapid, căci au tendința să fie defazate. Angajarea tot mai multor informaticieni experți în învățământul juridic și în guverne ar fi probabil o soluție bună. Ar trebui însă oare să apară sisteme de luare a deciziilor bazate pe IA

pentru votanți și legiuitori, urmate de robolegiuitori în toată puterea cuvântului?

Cum ar trebui modificate legile noastre pentru a reflecta progresul IA e o fascinantă temă de dezbatere. Un aspect se leagă de tensiunea dintre libertatea informației și caracterul ei privat. Partizanii libertății susțin că cu cât avem mai puțină confidențialitate, cu atât vor avea tribunalele mai multe probe, iar verdictele lor vor fi mai corecte. De pildă, dacă guvernul folosește dispozitivele electronice ale oricui pentru a înregistra unde e, ce tastează la calculator, ce spune și ce face, multe infracțiuni ar fi rapid lămurite, iar altele ar fi împiedicate. Apărătorii confidențialității răspund că nu-și doresc un stat orwellian care te supraveghează, și, chiar dacă și-ar dori, există riscul să se transforme într-o dictatură totalitară de proporții gigantice. În plus, tehnicile de învățare automată s-au perfecționat în analizarea datelor cerebrale ale scanărilor IRMF pentru a determina la ce se gândește un om și, în particular, pentru a determina dacă spune sau nu adevărul.³⁷ Dacă tehnologia de scanare cerebrală asistată de IA ar deveni ceva obișnuit în tribunale, stabilirea faptelor într-un proces ar fi mult simplificată și accelerată, iar verdictele ar fi mai corecte. Apărătorii confidențialității ar putea fi însă îngrijorați că asemenea sisteme ar comite din când în când erori, și mai ales că guvernul ne-ar spiona nelimitat. Guvernele care nu susțin libertatea de gândire ar putea folosi această tehnologie pentru a-i incrimina pe cei care au anumite credințe și opinii. Unde ai trasa *tu* granițele dintre justiție și confidențialitate și dintre protejarea societății și protejarea libertății personale? Indiferent unde ai trasa-o, se va deplasa ea oare treptat, dar inexorabil, către reducerea confidențialității pentru a compensa faptul că dovezile devin mai ușor de falsificat? De exemplu, odată ce IA poate genera imagini realiste cu tine comițând crime, ai vota pentru un sistem prin care guvernul urmărește în permanență pe toată lumea și care ți-ar oferi un alibi imbatabil, dacă ai avea nevoie de el?

Altă controversă captivantă se leagă de reglementarea IA sau, mai general, de felul în care factorii politici ar trebui să stimuleze

cercetările IA pentru a maximiza șansele unui rezultat benefic. Unii cercetători refuză toate formele de reglementare a dezvoltării IA, susținând că acestea ar întârzia inutil inovațiile necesare (de pildă, automobilele care, conducându-se singure, ar salva vieți) și ar face ca cercetarea IA de vârf să devină subterană și/sau ar transfera-o în alte țări, cu guverne mai permissive. La conferința din Puerto Rico, Elon Musk a susținut că din partea guvernelor avem nevoie de înțelegere, nu de supraveghere: experți ocupând poziții guvernamentale care să urmărească progresul IA și să-l îndrume. El a mai spus că reglementările guvernamentale pot să stimuleze uneori progresul, nu să-l înnăbușe: de exemplu, dacă standardele de siguranță guvernamentale pentru automobilele care se conduc singure pot duce la scăderea numărului accidentelor, atunci e mai puțin probabilă o reacție negativă din partea publicului, iar adoptarea noii tehnologii poate fi accelerată. Cele mai precaute companii IA ar favoriza prin urmare reglementările care îi obligă pe competitorii mai puțin scrupuloși să se conformeze standardelor înalte de siguranță.

O altă controversă juridică interesantă se referă la drepturile pe care le-am acorda mașinilor. Dacă automobilele care se conduc singure ar înjumătăți cele 32 000 de accidente rutiere mortale anuale din Statele Unite, poate că producătorii auto n-ar primi 16 000 de scrisori de mulțumire, ci 16 000 de trimiteri în judecată. Așadar, dacă un automobil care se conduce singur provoacă un accident, cine e răspunzător – pasagerii, proprietarul sau producătorul? Juristul David Vladeck propune un al patrulea răspuns: automobilul însuși! Mai exact, ideea e ca automobilelor care se conduc singure să li se permită (și să li se ceară) să aibă o asigurare auto. Astfel, modelele cu un „cazier“ excelent ar beneficia de mari reduceri, ajungându-se la sume mai mici decât cele disponibile pentru șoferii umani, în vreme ce modelele de calitate slabă ar avea polițe de asigurare care să le facă prohibitiv de scumpe.

Dar dacă mașinilor cum sunt automobilele li se permite să aibă polițe de asigurare, ar trebui să aibă ele voie și să dețină bani sau proprietăți? Dacă da, nu există nici un argument juridic care

să împiedice calculatoarele istețe să câștige bani la bursă și să-i folosească pentru a cumpăra servicii online. Odată ce un calculator începe să-i plătească pe oameni ca să lucreze pentru el, acesta poate realiza orice pot realiza oamenii. Dacă sistemele IA reușesc în cele din urmă să se priceapă mai bine decât oamenii la investiții (ceea ce se întâmplă deja în anumite domenii), s-ar putea ajunge la situația în care cea mai mare parte a economiei noastre să fie deținută și controlată de mașini. Asta ne dorim? Pare o perspectivă îndepărtată, însă majoritatea economiei noastre este de pe-acum deținută de o altă formă de entitate non-umană: corporațiile, care adesea sunt mai puternice decât orice persoană din cadrul lor și pot într-o anumită măsură să-și ia viața pe cont propriu.

Dacă acorzi mașinilor dreptul la proprietate, le-ai acorda și dreptul de a vota? În caz că da, ar trebui oare ca fiecare program de calculator să aibă un vot, deși pentru el e banal să-și facă bilioane de copii ale lui însuși în *cloud* (nor), dacă e suficient de bogat, hotărând astfel rezultatul tuturor alegerilor? În caz că nu, atunci pe ce bază morală defavorizăm mințile mașinilor în raport cu mințile umane? Are vreo importanță dacă mințile mașinilor sunt sau nu conștiente, în sensul că posedă o experiență subiectivă ca a noastră? Vom examina în detaliu problemele legate de controlul calculatorului asupra lumii noastre în capitolul următor, iar cele legate de conștiința mașinii în capitolul 8.

ARME

Din vremuri imemorabile omenirea a suferit din pricina foametei, a bolilor și a războaielor. Am văzut deja cum poate IA reduce foamea și bolile, dar ce putem spune despre războaie? Unii susțin că armele nucleare descurajează războaiele dintre țările care le dețin, fiindcă sunt atât de înspăimântătoare. Ar trebui așadar să permitem tuturor națiunilor să construiască arme și mai înspăimântătoare bazate pe IA, în speranța că nu vor mai exista niciodată războaie? Dacă nu ești convins de acest argument și crezi

că războaiele viitoare sunt inevitabile, ce părere ai de folosirea IA pentru a face ca aceste războaie să fie mai umane? Dacă războaiele se reduc la mașini luptând împotriva mașinilor, atunci nu mai sunt uciși soldați sau civili umani. În plus, viitoarele drone acționate de IA și alte sisteme de arme autonome (*autonomous weapon systems* – AWS; numite de oponenții lor și „roboți ucigași”) este de sperat să fie mai corecte și mai raționale decât soldații umani: echipate cu senzori supraumani și netemându-se că vor fi ucise, ar rămâne reci, calculate și echilibrate chiar și în toiul bătăliei, fiind mai puțin probabil să ucidă în mod accidental civili.



Figura 3.4 Dacă dronele militare de azi (cum e această MQ-1 Predator a Forțelor Aeriene ale Statelor Unite) sunt controlate de la distanță de oameni, viitoarele drone acționate de IA ar putea lăsa oamenii pe dinafară, folosind un algoritm care să hotărască ce țintă trebuie atinsă.

Un om în circuitul deciziei

Dar dacă sistemele automate conțin erori, sunt derutante sau nu se comportă cum ne așteptăm? Sistemul american Phalanx pentru crucișătoarele din clasa Aegis detectează, urmărește și atacă rachete sau avioane amenințătoare. Vincennes era un crucișător american botezat Robocruiser datorită automatizărilor cu care era echipat, iar pe 3 iulie 1988, în toiul unei încăierări cu canonierele iraniene din timpul războiului dintre Iran și Irak, radarul

lui a detectat un avion care se apropia. Căpitanul William Rodgers III a dedus că erau atacați în picaj de un avion de vânătoare iranian F-14 și a dat sistemului Aegis aprobarea să deschidă focul. Nu și-a dat seama însă că a doborât zborul Iran Air 655, un avion iranian civil de pasageri, ucigându-i pe toți cei 290 de oameni aflați la bordul lui și provocând indignarea întregii lumi. Investigațiile ulterioare au scos la iveală că de vină era interfața om-mașină, care nu arăta automat ce puncte de pe ecranul radarului erau avioane civile (zborul 655 își urma cursul zilnic regulat, transponderul lui de aeronavă civilă fiind în funcțiune) și ce puncte coborau (ca pentru atac) sau urcau (așa cum se întâmpla cu zborul 655, după decolarea din Teheran). Când sistemului automat i s-au cerut informații despre misteriosul avion, a raportat că „se află în coborâre”, fiindcă asta făcea un alt avion, căruia din greșeală îi fusese realocat un număr folosit de forțele navale pentru a urmări avioanele: ceea ce cobora era un avion american care patrula departe, în Golful Oman.

În acest exemplu, decizia finală a fost luată de un om aflat în circuit, care, sub presiunea timpului, a avut prea multă încredere în ce-i spunea sistemul automat. Conform responsabililor militari din lumea întreagă, toate sistemele de armament desfășurate deocamdată includ un om în circuitul deciziei, cu excepția capcanelor primitive de tipul minelor terestre. Sunt însă concepute acum arme cu adevărat autonome, care selectează și atacă ținte pe cont propriu. Din punct de vedere militar e tentant să scoți toți oamenii din circuitul deciziei pentru a spori viteza: într-o luptă dintre o dronă complet autonomă, care poate răspunde instantaneu, și o dronă care reacționează mai lent, fiindcă e controlată de oameni de la mare distanță, care credeți că ar învinge?

Pe de altă parte însă, au existat situații în care am fost extrem de norocoși că a existat un om în circuitul deciziei. Pe 27 octombrie 1962, în timpul crizei rachetelor din Cuba, unsprezece distrugătoare și portavionul Randolph încolțiseră un submarin sovietic B-59 în apropierea coastei cubaneze, în apele internaționale din afara zonei americane de „carantină”. Ceea ce nu se știa

era că temperatura la bordul lui ajunsese la 45°C, pentru că bateriile submarinului se descărcaseră, iar aerul condiționat nu mai funcționa. Intoxicați cu dioxid de carbon, mulți membri ai echipajului leșinaseră. Trecuseră câteva zile de când nu mai comunicaseră cu Moscova și nu știau dacă nu cumva al Treilea Război Mondial începuse deja. Americanii au lansat atunci mici grenade antisubmarin, și au spus Moscovei că nu voiau decât să oblige submarinul să iasă la suprafață și să plece. „Credeam că, gata, s-a sfârșit”, avea să-și amintească V.P. Orlov, membru al echipajului sovietic. „Simțeau că stai într-un butoi de metal pe care cineva îl izbește continuu cu un baros.” Ceea ce iarăși nu știau americanii era că B-59 avea o torpilă nucleară pe care o putea lansa fără să aștepte aprobarea Moscovei. Și, într-adevăr, căpitanul Savițki s-a hotărât să lanseze torpila nucleară. Valentin Grigorievici, ofițerul responsabil de torpilă, a exclamat: „O să murim, dar o să-i scufundăm pe ei, pe toți, n-o să facem de râs flota noastră militară!” Din fericire, decizia de lansare trebuia avizată de trei ofițeri, iar unul dintre ei, Vasili Arhipov, a spus nu. E trist că foarte puțini au auzit de Arhipov, deși hotărârea lui a împiedicat pesemne declanșarea celui de-al Treilea Război Mondial, și a fost cea mai prețioasă contribuție individuală pentru salvarea omenirii din istoria modernă.³⁸ E cutremurător să te gândești ce s-ar fi întâmplat dacă B-59 ar fi fost un submarin autonom controlat de IA, fără oameni în circuitul deciziei.

Două decenii mai târziu, pe 9 septembrie 1983, tensiunea dintre superputeri crescuse din nou: Uniunea Sovietică fusese recent numită „imperiul răului” de președintele american Ronald Reagan, iar în săptămâna precedentă doborâse un avion de pasageri al companiei Korean Airlines care se rătăcise în spațiul ei aerian, ucigând 269 de oameni, între care un congressman american. Un sistem sovietic automat de avertizare timpurie a raportat că Statele Unite lansaseră asupra Uniunii Sovietice cinci rachete nucleare, lăsând la dispoziția ofițerului Stanislav Petrov doar câteva minute pentru a hotărî dacă era o alarmă falsă. Satelitul funcționa corect, deci, urmând protocolul, trebuia să anunțe un atac nuclear.

Urmându-și însă instinctul și socotind că era improbabil ca americanii să atace cu numai cinci rachete, a raportat comandanților săi că era o alarmă falsă, fără să știe că acesta era adevărul. S-a aflat mai târziu că satelitul luase reflexiile Soarelui pe suprafața superioară a norilor drept flăcări ale unor motoare de rachetă.³⁹ Mă întreb ce s-ar fi întâmplat dacă Petrov ar fi fost înlocuit de un sistem IA care să urmeze întocmai protocolul.

Următoarea cursă a înarmărilor?

După cum bănuiai deja, mă îngrijorează foarte mult sistemele de arme autonome. Dar n-am ajuns încă la principala mea temere: punctul final al unei curse a înarmărilor cu armament IA. În iulie 2015 mi-am exprimat îngrijorarea în următoarea scrisoare deschisă, alături de Stuart Russell, având susținerea colegilor mei de la Institutul pentru Viitorul Vieții:⁴⁰

ARMELE AUTONOME

Scrisoare deschisă din partea cercetătorilor în domeniul inteligenței artificiale și roboticii

Armele autonome selectează și atacă ținte fără intervenție umană. Ele ar putea include, de exemplu, quadcoptere* înarmate care pot depista și elimina oameni îndeplinind anumite criterii prestabilite, dar nu includ rachete de croazieră sau drone pilotate de la distanță, la care oamenii iau toate deciziile de lovire a țintei. Tehnologia IA a ajuns într-un punct în care punerea în funcțiune a unor asemenea sisteme e practic, chiar dacă nu și legal, posibilă în următorii ani, nu peste decenii, iar miza este enormă: armele autonome reprezintă a treia revoluție în domeniul militar, după praful de pușcă și armele nucleare.

S-au adus multe argumente pro și contra armelor autonome – de pildă, că înlocuirea soldaților umani cu mașini e bună prin reducerea numărului victimelor de partea celui care le deține,

* Elicoptere ridicate și propulsate de patru rotoare. (N. t.)

dar rea prin scăderea pragului în declanșarea bătăliilor. Întrebarea-cheie pentru omenire este acum dacă să înceapă o cursă globală a înarmărilor IA sau s-o împiedice. Dacă o mare putere militară pornește dezvoltarea armei IA, atunci o cursă a înarmărilor e practic inevitabilă, iar punctul final al acestei traiectorii tehnologice este evident: armele autonome vor deveni kalașnikovurile de mâine. Spre deosebire de armele nucleare, ele nu necesită materii prime scumpe sau greu de obținut, ca atare vor deveni omniprezente, iar pentru marile puteri militare producția lor în masă va fi ieftină. Va fi doar o chestiune de timp până să apară pe piața neagră și în mâinile teroriștilor, ale dictatorilor care vor să controleze mai bine masele, ale războinicilor dornici de purificare etnică etc. Armele autonome sunt ideale pentru sarcini precum asasinatele, destabilizarea națiunilor, supunerea populațiilor și uciderea selectivă a unui anumit grup etnic. Credem așadar că o cursă a înarmărilor IA nu va fi în folosul omenirii. Există multe moduri prin care IA poate face câmpurile de bătălie mai sigure pentru oameni, în special pentru civili, fără a crea noi instrumente de ucidere a oamenilor.

Așa cum majoritatea chimiștilor și biologilor n-au interesul să construiască arme chimice sau biologice, majoritatea cercetătorilor IA n-au interesul să construiască arme IA și nu vor ca alții să întineze domeniul lor făcând acest lucru, provocând pesemne o puternică reacție din partea publicului împotriva IA, care să diminueze viitoarele ei beneficii sociale. Într-adevăr, chimiștii și biologii au sprijinit masiv înțelegerile internaționale care au reușit să interzică armele chimice și biologice, la fel cum majoritatea fizicienilor au sprijinit tratatele interzicând plasarea în spațiu a armelor nucleare și a armelor laser orbitoare.

Pentru a face mai greu de respins îngrijorările noastre, pe motiv că sunt ale unor ecologiști pacifiști, am vrut ca scrisoarea să fie semnată de cât mai mulți cercetători IA și roboticieni de vârf. Campania Internațională pentru Controlul Armelor Robotice adunase

sute de semnături care cereau interzicerea roboților ucigași, iar eu speram ca noi să ne descurcăm și mai bine. Știam că organizațiile profesionale n-ar fi dispuse să ne ofere listele cu adresele de e-mail ale membrilor lor pentru un scop care putea fi considerat politic, așa că am încropit liste cu numele cercetătorilor și instituții din documente online și am pus sarcina găsirii adreselor de e-mail pe seama platformei Amazon MTurk. Majoritatea cercetătorilor au adresele de e-mail trecute în site-urile universităților lor, iar după 24 de ore și după ce am plătit 54 de dolari, eram mândrul posesor al unei liste cu adresele a sute de cercetători IA suficient de reputați pentru a fi aleși membri ai Asociației pentru Progresul Inteligenței Artificiale (Association for the Advancement of Artificial Intelligence – AAAI). Unul dintre ei era profesorul australian Toby Walsh, care a trimis mesaje tuturor celor de pe listă și a difuzat campania noastră. Participanții la MTurk din lumea întreagă au produs noi liste pentru Toby, iar curând peste 3 000 de cercetători în IA și robotică semnaseră scrisoarea noastră deschisă, între care șase foști președinți AAAI și lideri ai industriei IA de la Google, Facebook, Microsoft și Tesla. O armată de voluntari FLI a lucrat neobosit pentru a valida listele de semnături, eliminând nume trecute în bătaie de joc ca Bill Clinton și Sarah Connor. S-au mai adăugat alte 17 000 de semnături, între care cea a lui Stephen Hawking, iar după ce Toby a organizat o conferință de presă pe această temă la Conferința Internațională Comună a Inteligenței Artificiale, scrisoarea a intrat în atenția marilor agenții de presă din lumea întreagă.

Pentru că biologii și chimiștii au luat la un moment dat atitudine, domeniile lor sunt acum cunoscute în primul rând pentru crearea medicamentelor și materialelor folositoare, nu a armelor biologice și chimice. Acum s-au exprimat și comunitățile din IA și robotică: semnatarii scrisorii voiau de asemenea ca domeniile lor să fie cunoscute pentru făurirea unui viitor mai bun, nu a unor noi moduri de a ucide oameni. Dar principala utilizare a IA va fi civilă sau militară? Deși în acest capitol am vorbit mai mult despre prima, s-ar putea ca în curând să cheltuim mai mulți bani pentru

cea de-a doua – mai ales dacă se declanșează o cursă a înarmărilor IA. Investițiile în IA din domeniul civil au depășit în 2016 un miliard de dolari, dar Pentagonul a cerut pentru anul 2017 un buget de 12–15 miliarde de dolari destinați proiectelor legate de IA, iar China și Rusia n-au rămas pesemne indiferente la această declarație a lui Robert Work, ministrul adjunct al Apărării: „Vreau ca rivalii noștri să se întrebe se s-o fi ascunzând în spatele cortinei negre.”⁴¹

Ar trebui semnat un tratat internațional?

Deși există acum presiuni internaționale pentru negocieri privind interzicerea unei anumite forme de roboți ucigași, nu-i limpede ce se va întâmpla, și s-a declanșat o dezbatere în legătură cu ce anume ar *trebui* să se întâmple. Multe figuri importante din domeniu sunt de acord că e nevoie ca puterile mondiale să elaboreze un fel de reglementări internaționale pentru a îndruma cercetările și utilizarea AWS, dar nu cad de acord cu privire la ce anume trebuie interzis și cum poate fi impusă o interdicție. De pildă, ar trebui interzise numai armele autonome letale, sau și cele care produc răni grave, prin sângerare? Să interzicem proiectarea, producerea sau deținerea? Interdicția ar trebui aplicată tuturor sistemelor de arme autonome, sau, așa cum se spune în scrisoarea noastră, doar celor ofensive, acceptându-se sisteme defensive cum sunt armele autonome antiaeriene și antirachetă? În acest din urmă caz, AWS trebuie socotite defensive chiar dacă sunt ușor de deplasat în teritoriul inamic? Și cum poți aplica un tratat, când majoritatea componentelor armelor autonome au și o utilizare civilă? De exemplu, nu-i mare diferență între o dronă încărcată cu pachete de la Amazon și una încărcată cu bombe.

Unii participanți la dezbateri au afirmat că elaborarea unui tratat AWS eficient e extrem de dificilă și nici n-are rost să încerci. Pe de altă parte, anunțând programul spațial care avea ca obiectiv Luna, John F. Kennedy a spus că merită să încerci să faci lucruri dificile atunci când succesul va fi important pentru viitorul omenirii. Apoi, mulți experți au susținut că interzicerea armelor

biologice și chimice s-a dovedit benefică, chiar dacă impunerea ei a fost problematică, iar unii au trișat, fiindcă interzicerea a provocat o gravă stigmatizare care a limitat folosirea acelor arme.

În 2016 m-am întâlnit la un dîneu cu Henry Kissinger, și am avut ocazia să-l întreb despre rolul pe care l-a jucat în interzicerea armelor biologice. Mi-a explicat că, pe când era consilier pentru securitatea națională, l-a convins pe președintele Nixon că interdicția ar fi în folosul securității americane. M-au impresionat mintea ascuțită și memoria unui om de 92 de ani, și m-a fascinat perspectiva lui de politician. Statele Unite având deja statutul de supraputere grație forțelor militare convenționale și nucleare, aveau mai curând de pierdut decât de câștigat dintr-o cursă mondială a înarmărilor biologice cu rezultat incert. Altfel spus, dacă te afli deja în frunte, e util să urmezi principiul „Dacă nu-i stricat, nu-l repara”. Stuart Russell s-a alăturat conversației noastre, și am discutat despre aplicarea întocmai a aceluiași argument în cazul armelor autonome letale: nu supraputerile ar avea de câștigat dintr-o cursă a înarmărilor, ci micile state lipsite de scrupule și actorii nonstatali precum teroriștii, care ar avea acces la arme prin piața neagră, odată ce ele vor fi fost construite.

Trecându-se la producția în masă, micile drone ucigașe echipate cu IA e probabil să coste cu puțin mai mult decât un telefon inteligent. Indiferent dacă un terorist vrea să asasineze un politician ori un bărbat părăsit caută să se răzbune pe fosta lui iubită, e de-ajuns ca ei să încarce fotografia țintei și adresa ei în drona ucigașă: aceasta va zbura spre destinație, va identifica și elimina persoana, iar apoi se va autodistruge, pentru ca nimeni să nu știe cine e răspunzător. Sau, pentru cei care urmăresc purificarea etnică, aceasta poate fi lesne programată să ucidă doar oameni cu o anumită culoare a pielii ori de o anumită etnie. Stuart prevede că, cu cât vor deveni mai deștepte asemenea arme, cu atât va fi nevoie de mai puține materiale, muniții și bani per crimă. De pildă, se teme de drone de dimensiunea unei albine care vor ucide folosind muniție minimă împușcând în ochi, care e suficient de moale pentru a permite unui mic proiectil să pătrundă în creier. Sau s-ar putea

agăța de cap cu niște gheare metalice, pentru a străpunge apoi craniul cu o încărcătură minusculă. Dacă un milion de asemenea drone ucigașe pot fi lansate din remorca unui singur camion, avem o înspăimântătoare armă de distrugere în masă de un tip cu totul nou, care ucide selectiv doar o anumită categorie de oameni.

Un contraargument este că putem elimina aceste griji făcând ca roboții ucigași să fie „etici” – de pildă, să ucidă numai soldații inamici. Dar dacă ni se pare greu să aplicăm o interdicție, ar fi oare mai ușor să impunem condiția ca armele autonome să fie 100% etice decât să impunem ca ele să nu fie produse? Și putem afirma că roboții ar asculta cu mai multă strictețe regulile războiului decât soldații bine antrenați ai națiunilor civilizate, iar pe de altă parte că dictatorii și grupările teroriste respectă atât de strict regulile războiului, încât n-ar alege niciodată să folosească roboții într-un mod care violează aceste reguli?

Războiul cibernetic

Alt aspect militar interesant al IA este că aceasta ți-ar putea permite să-ți ataci dușmanul chiar și fără să construiești vreo armă, prin război cibernetic. Ca un mic preludiu a ceea ce viitorul ne-ar putea aduce, viermele Stuxnet, despre care se bănuiește că a fost creat de americani și israelieni, a infectat centrifugele din programul nuclear iranian de îmbogățire izotopică și le-a dezmembrat. Cu cât o societate e mai automatizată și cu cât devine mai puternic atacul IA, cu atât mai devastator poate fi războiul cibernetic. Dacă poți virusa și distruge automobilele care se conduc singure ale inamicului, avioanele lui autopilotate, reactoarele nucleare, roboții industriali, sistemele financiar și de comunicații, rețelele electrice, atunci îi poți efectiv distruge economia și paraliza apărarea. Dacă îi poți virusa și sistemele militare, cu atât mai bine.

Am început acest capitol arătând cât de benefică poate fi pentru omenire, în viitorul apropiat, IA – cu condiția să devină robustă și imună la atacurile informatice. Deși IA însăși poate fi folosită pentru a face sistemele IA mai robuste, sprijinind astfel apărarea în războiul cibernetic, ea poate sprijini și agresiunea. Garantarea

faptului că apărarea predomină trebuie să fie unul dintre obiectivele pe termen scurt ale dezvoltării IA – altminteri întreaga tehnologie pe care o construim se poate întoarce împotriva noastră!

JOBURI ȘI SALARII

Până acum acest capitol s-a referit în primul rând la impactul IA asupra noastră în calitate de *consumatori*, oferindu-ne noi produse și servicii la prețuri accesibile. Dar cum ne va influența în calitate de *muncitori*, transformând piața muncii? Dacă găsim o cale de a ne spori prosperitatea prin automatizare, fără a-i priva pe oameni de venituri sau scopuri, atunci ne putem făuri un viitor fantastic, cu un timp liber și o bogăție nemaiîntâlnite pentru toată lumea. Puțini s-au gândit la acest aspect mai îndelungat și mai profund decât economistul Erik Brynjolfsson, colegul meu de la MIT. Deși e mereu îngrijit și impecabil îmbrăcat, ascendența sa islandeză mă face să-mi închipui că abia de curând și-a ras barba roșie de viking și și-a tuns chica pentru a se alătura școlii noastre de afaceri. Fără îndoială însă nu s-a debarasat și de ideile lui îndrăznețe, iar viziunea sa optimistă asupra pieței muncii a numit-o „Atena Digitală”. Motivul principal pentru care cetățenii atenieni din Antichitate au dus vieți confortabile, în care se puteau bucura de democrație, artă și jocuri, a fost că aveau sclavi care să facă mare parte din muncă. De ce să nu înlocuim deci sclavii cu roboți echipați cu IA, creând o utopie digitală de care toată lumea să se bucure? Nu numai că ar elimina stresul și corvoada și ar produce tot ce ne dorim azi, dar economia condusă de IA, în viziunea lui Erik, ne-ar oferi și o mulțime de noi produse și servicii pe care consumatorii de azi nici măcar nu și-au dat seama că și le doresc.

Tehnologie și inegalitate

Putem ajunge la Atena Digitală a lui Erik dacă salariul pe oră al fiecăruia continuă să crească an de an, așa încât cei care vor mai mult timp liber să poată treptat munci mai puțin, îmbunătățindu-și

constant nivelul de trai. Figura 3.5 arată că tocmai asta s-a întâmplat în Statele Unite de la Al Doilea Război Mondial până pe la mijlocul anilor '70: deși a existat o inegalitate a veniturilor, dimensiunea totală a tortului a crescut așa încât aproape toți au primit o felie mai mare. Apoi însă, după cum recunoaște și Erik, ceva s-a schimbat: figura 3.5 arată că deși economia a continuat să crească, sporind venitul mediu, câștigurile din ultimele patru decenii s-au îndreptat către cei bogați, în special către cei mai bogați 1%, iar la cei mai săraci 90% veniturile au stagnat. Accentuarea inegalității reiese și mai clar dacă ne uităm nu la venit, ci la nivelul bunăstării. Pentru cele mai sărace 90% dintre familiile americane, activul net mediu a fost de circa 85 000 de dolari în 2012 – același ca în urmă cu douăzeci de ani –, în vreme ce pentru cele mai bogate 1% s-a dublat în aceeași perioadă, ajungând la 14 milioane de dolari (luând în calcul inflația).⁴² Pe plan mondial diferențele sunt și mai mari. În 2013, averea însumată a jumătății mai sărace din populația lumii (peste 3,6 miliarde de oameni) este egală cu averea celor mai bogați opt oameni.⁴³ La conferința noastră din Puerto Rico, Erik a afirmat că progresul inteligenței artificiale și automatizării va face să crească tortul economiei, dar nici o lege economică nu spune că toată lumea, sau măcar majoritatea, va avea de câștigat.

Deși economiștii cad de acord asupra faptului că inegalitatea se accentuează, există o interesantă controversă privind cauzele fenomenului și tendințele viitoare. Cei cu orientări politice de stânga susțin că principalele cauze sunt globalizarea și/sau politicile economice precum reducerile fiscale pentru cei bogați. Erik Brynjolfsson și colaboratorul său de la MIT Andrew McAfee susțin însă că alta e principala cauză: tehnologia.⁴⁴ Mai precis, tehnologia digitală acționează în trei moduri asupra inegalității.

În primul rând, prin înlocuirea vechilor joburi cu unele care necesită mai multe aptitudini, tehnologia i-a recompensat pe cei educați: de la mijlocul anilor '70, salariile absolvenților de facultate au crescut cu circa 25%, iar salariile celor care au abandonat școala au scăzut cu circa 30%.⁴⁵

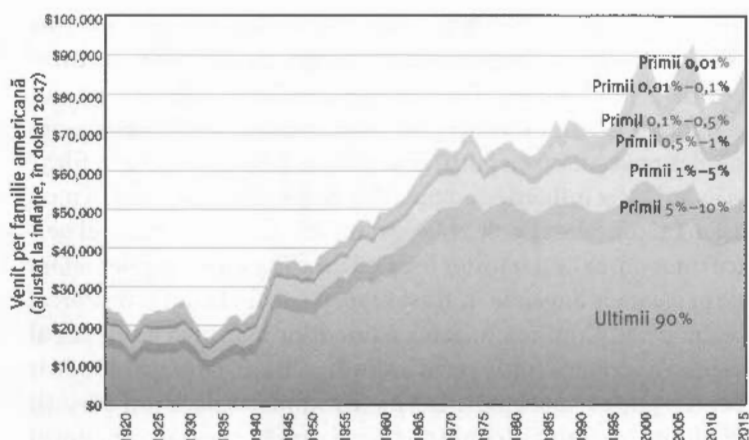


Figura 3.5 Creșterea venitului mediu în ultimul secol și partea din acesta care a revenit diferitelor grupuri. Înainte de anii '70, bogați și săraci au prosperat în același ritm, dar apoi majoritatea câștigurilor s-au îndreptat către primii 1%, iar ultimii 90% n-au câștigat în medie aproape nimic.⁴⁶ Sumele au fost corectate în raport cu inflația și reprezintă dolari din 2017.

În al doilea rând, începând din anul 2000, o parte tot mai mare din venitul corporațiilor s-a îndreptat către cei care dețin companiile, nu către cei care muncesc în ele – iar atâta timp cât automatizarea continuă, este de așteptat ca aceia care dețin mașinile să capete o felie tot mai mare din tort. Raportul dintre capital și muncă poate juca un rol foarte important pentru economia digitală în ascensiune, pe care Nicholas Negroponte, un profet al tehnologiei, o definește prin mișcarea biților, nu a atomilor. Acum când totul, de la cărți la filme și la instrumentele pentru întocmirea declarațiilor fiscale, a devenit digital, copiile pot fi vândute pretutindeni în lume cu costuri practic nule, fără angajarea unui personal suplimentar. Aceasta permite ca majoritatea profitului să revină investitorilor, nu muncitorilor, și explică de ce profitul însumat al primelor trei companii din Detroit (General Motors, Ford și Chrysler) în 1990 a fost aproape identic cu profitul însumat al primelor trei companii din Silicon Valley (Google, Apple, Facebook) în 2014, acestea din urmă având de nouă ori mai puțini angajați și valorând de treizeci de ori mai mult la bursă.⁴⁷

În al treilea rând, Erik și colaboratorii săi susțin că economia digitală favorizează adesea superstarurile în detrimentul tuturor celorlalți. J.K. Rowling, autoarea lui Harry Potter, a devenit primul scriitor miliardar, și s-a îmbogățit mult mai mult decât Shakespeare fiindcă poveștile ei au putut fi transmise sub formă de text, filme și jocuri către miliarde de oameni la un preț foarte scăzut. În mod asemănător, Scott Cook a câștigat un miliard cu software-ul pentru întocmirea declarațiilor fiscale TurboTax, care, spre deosebire de persoanele angajate în acest scop, poate fi vândut și descărcat pe internet. Cum majoritatea oamenilor n-ar prea cumpăra al zecelea asemenea software în ordinea calității, pe piață e loc doar pentru puține superstaruri. Asta înseamnă că dacă toți părinții din lume își sfătuiesc copiii să devină următoarea sau următorul J.K. Rowling, Gisele Bundchen, Matt Damon, Cristiano Ronaldo sau Elon Musk, aproape nici unul dintre copii nu va găsi că e o strategie viabilă pentru cariera de urmat.

Sfaturi profesionale pentru copii

Ce carieră *ar trebui* să-i sfătuim pe copiii noștri să urmeze? Eu i-am încurajat pe ai mei să caute profesii în care mașinile deocamdată nu se descurcă, fiind deci improbabil să se automatizeze în viitorul apropiat. Previziuni recente pentru momentul când diferite joburi vor fi preluate de mașini scot în evidență câteva întrebări pe care e util să ți le pui înainte de a începe educația pentru a urma o anumită carieră.⁴⁸ De exemplu:

- Presupune interacțiunea cu oamenii și folosirea inteligenței sociale?
- Implică oare creativitatea și găsirea unor soluții ingenioase?
- Presupune munca într-un domeniu imprevizibil?

Cu cât poți răspunde prin da la mai multe dintre aceste întrebări, cu atât e probabil să fie mai bună alegerea carierei. Asta înseamnă că profesii relativ sigure sunt cele de profesor, asistentă medicală, doctor, dentist, om de știință, antreprenor, inginer, avocat, asistent social, cleric, artist, coafeză sau kinetoterapeut.

Dimpotrivă, joburi care presupun acțiuni repetitive sau structurate, în condiții predictibile, este probabil să devină nu peste mult timp automatizate. Computerele și roboții industriali au preluat de mult cele mai simple asemenea joburi, iar tehnologia în curs de perfecționare elimină multe altele, de la agent de vânzări la lucrător în depozit, casier, feroviar, brutar și bucătar necalificat.⁴⁹ Șoferii de camioane, autobuze, taxiuri și mașini Uber/Lyft e probabil să urmeze în curând. Există numeroase alte profesii (între care asistent juridic, analist de credite și contabil) care, deși nu sunt încă amenințate cu dispariția completă, își văd deja automatizate majoritatea sarcinilor, prin urmare necesită mai puțini oameni.

Dar evitarea automatizării nu e singura provocare legată de carieră. În această epocă digitală globală, a dori să devii scriitor profesionist, regizor de film, actor, atlet sau creator de modă e riscant din alt motiv: deși oamenii din aceste profesii nu vor fi concurați serios de mașini în viitorul apropiat, ei vor fi tot mai puternic concurați de alți oameni de pretutindeni, conform teoriei superstarului menționată mai sus, iar foarte puțini vor reuși.

În multe cazuri, e imprecis să dai sfaturi la nivelul unui întreg domeniu: numeroase joburi, fără a fi complet eliminate, vor avea multe dintre sarcinile lor automatizate. De pildă, dacă vrei să studiezi medicina, nu deveni acel radiolog care analizează imaginile și e înlocuit de calculatorul IBM Watson, ci doctorul care prescrie analiza radiologică, discută rezultatele cu pacientul și hotărăște tratamentul. Dacă vrei să intri în finanțe, nu deveni acel analist cantitativ care aplică algoritmi asupra datelor și e înlocuit de un software, ci administrator de fonduri care folosește rezultatele analizei cantitative pentru a lua decizii strategice de investiții. Dacă vrei să intri în domeniul juridic, nu deveni acel asistent care parcurge mii de acte pentru faza de documentare și e eliminat prin automatizare, ci avocatul care își sfătuiește clientul și pledează la tribunal.

Am examinat deocamdată ce pot face indivizii pentru a-și maximiza succesul pe piața muncii în epoca IA. Dar ce pot face

guvernele pentru a veni în sprijinul forței de muncă? De pildă, ce sistem de învățământ oferă cea mai bună pregătire pentru o piață a muncii în care IA evoluează rapid. Este în continuare actualul nostru model cu două decenii de învățământ, urmate de patru decenii de muncă specializată? Sau e de preferat ca oamenii să lucreze câțiva ani, să se întoarcă la studii pentru un an, apoi să lucreze iar câțiva ani?⁵⁰ Sau continuarea educației (eventual online) să fie regula oricărui job?

Ce politici economice sunt cele mai utile pentru crearea unor joburi noi profitabile? Andrew McAfee susține că există destule asemenea politici, între care investiții masive în cercetare, educație și infrastructură, facilitarea migrației și stimularea antreprenoriatului. El crede că „playbook-ul Econ 101 oferă sfaturi clare, dar nu e urmat“, cel puțin în Stele Unite.⁵¹

Vor deveni oamenii în cele din urmă neangajabili?

Ce se va întâmpla dacă IA continuă să se perfecționeze, automatizând tot mai multe joburi? Mulți oameni rămân optimiști, afirmând că joburile automatizate vor fi înlocuite de unele noi, mai bune. În fond, așa s-a întâmplat mereu până acum, de când, în timpul Revoluției Industriale, pe ludiți i-a îngrijorat șomajul tehnologic.

Alții sunt însă pesimiști și susțin că de data asta e altfel, iar tot mai mulți oameni vor deveni nu doar șomeri, ci neangajabili.⁵² Pesimiștii afirmă că piața liberă stabilește salariile pe baza cererii și ofertei, și că oferta sporită de muncă automată ieftină va duce în cele din urmă la scăderea salariilor oamenilor mult sub costul vieții. Din moment ce salariul pe piață pentru un job e dat de cât costă pe oră oricine sau orice efectuează cel mai ieftin acea muncă, salariile au scăzut de fiecare dată când a fost posibil să externalizezi o anumită ocupație către o țară mai săracă sau o mașină ieftină. În timpul Revoluției Industriale, am început să ne gândim cum să înlocuim mușchii noștri cu mașini, iar oamenii s-au orientat spre joburi mai bine plătite, în care își foloseau mai mult mintea. „Gulerele albastre“ au fost înlocuite de „gulerele albe“. Acum ne gândim cum să înlocuim mințile noastre cu mașini. Dacă vom reuși, atunci ce joburi rămân pentru noi?

Unii optimiști susțin că, după joburile fizice și mentale, următorul boom va fi în joburile creative, dar pentru pesimiști creativitatea nu-i decât un proces mental, așa încât și ea va fi deprinsă în cele din urmă de IA. Alți optimiști speră ca următorul boom să fie în noi profesii apărute grație tehnologiei, pe care nici nu ni le-am închipuit până acum. Cine și-ar fi imaginat în timpul Revoluției Industriale că urmașii lui vor lucra într-o bună zi ca web designeri sau ca șoferi Uber? Pentru pesimiști însă, acestea sunt iluzii deșarte, nesusținute de datele empirice. Ei subliniază că același raționament l-am fi putut face cu un secol în urmă, înaintea revoluției informatice, și am fi prevăzut că majoritatea profesiilor de azi vor fi fiind unele noi și nebanuite, apărute grație tehnologiei. Această predicție ar fi fost eronată, după cum se vede

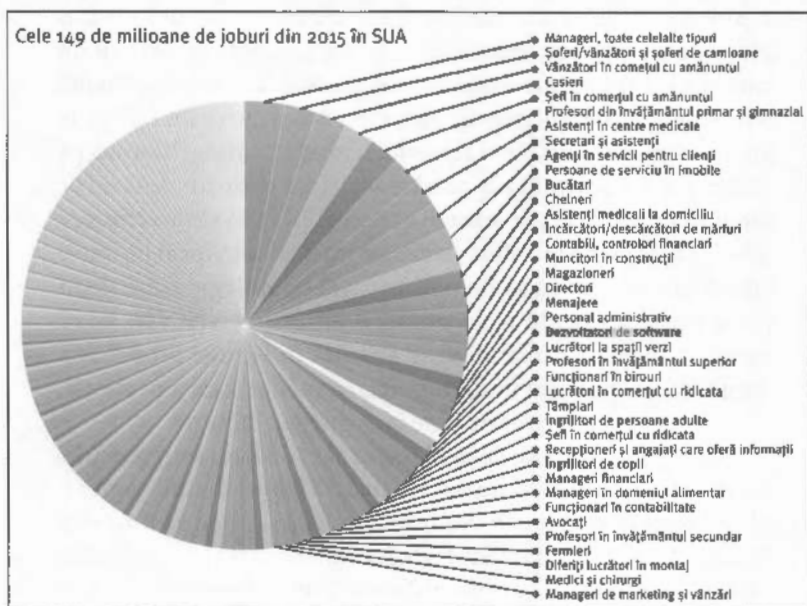


Figura 3.6 Graficul reprezintă ocupațiile celor 149 de milioane de americani care aveau un job în 2015, cu cele 535 de categorii de joburi conform Departamentului de Statistică a Muncii, clasificate în ordinea descrescătoare a numărului de lucrători.⁵³ Sunt marcate toate ocupațiile cu peste un milion de lucrători. Până la poziția 21 nu există noi ocupații create de tehnologia informaticii. Figura e inspirată de o analiză a lui Federico Pistono.⁵⁴

în figura 3.6: marea majoritate a ocupațiilor de azi sunt cele care existau deja cu un secol în urmă, iar dacă le clasificăm după numărul de joburi pe care le oferă, ajungem la poziția 21 de pe listă pentru a găsi o nouă ocupație: dezvoltatori de software, care reprezintă mai puțin de 1% din piața muncii în Statele Unite.

Înțelegem mai bine ce se întâmplă dacă ne amintim de figura 2.2 din capitolul 2, care prezintă peisajul inteligenței umane, altitudinea exprimând cât de dificil le este mașinilor să execute diferite sarcini, iar nivelul în creștere al apei indicând ce pot face acum mașinile. Tendința principală a pieței muncii nu e de orientare către profesii cu totul noi. Ne înghesuim în acele zone de uscat din figura 2.2 care n-au fost încă inundate de marea tehnologiei! Figura 3.6 arată că ele nu formează o singură insulă, ci un arhipelag complex, cu insulițe și atoluri corespunzând tuturor acelor lucruri pe care mașinile nu le pot face deocamdată cu un cost la fel de mic ca oamenii. E vorba nu doar de profesii de înaltă tehnologie, ca dezvoltator de software, dar și de o gamă largă de joburi implicând un nivel tehnologic scăzut, care se bazează pe dexteritatea și aptitudinile noastre sociale superioare, de la cel de kinetoterapeut la cel de actor. E oare posibil ca IA să ne eclipseze atât de rapid în îndeplinirea sarcinilor intelectuale, încât ultimele joburi care ne rămân să fie cele de nivel tehnologic scăzut? Un prieten mi-a zis în glumă că probabil ultima meserie care ne rămâne va fi chiar prima meserie apărută în istorie: prostituția. Apoi însă i-a spus asta unui robotician japonez, care a protestat: „Nu, roboții se pricep foarte bine la asemenea lucruri!”

Pesimiștii afirmă că deznodământul e evident: întregul arhipelag va fi scufundat, și nu vor mai exista joburi în care oamenii să muncească mai ieftin decât mașinile. În cartea sa din 2007 *Farewell to Alms*, economistul Gregory Clark arată că putem afla câte ceva despre viitorul nostru dintr-o comparație cu cabalinele. Închipuți-vă doi cai asistând pe la 1900 la începuturile automobilului și meditănd asupra viitorului lor.

„Mă îngrijorează șomajul tehnologic.”

„Hai, nu fi ludic. Strămoșii noștri au spus același lucru atunci când motoarele cu aburi ne-au luat joburile din industrie și

trenurile ne-au luat jobul de a trage poștalioanele. Avem azi mai multe joburi ca niciodată, și sunt și mai bune: prefer să trag o trăsură ușoară prin oraș decât să merg ziua întreagă în cerc ca să acționez o pompă stupidă dintr-o mină.“

„Dar dacă motoarele cu ardere internă se dezvoltă?”

„Sunt sigur că vor fi joburi noi pentru cai, pe care încă nu ni le imaginăm. Așa s-a întâmplat de fiecare dată până acum, la fel ca atunci când s-au inventat roata și plugul.“

Din păcate, acele joburi încă neimagineate pentru cai n-au mai apărut niciodată. Caii de care nu era nevoie au fost sacrificați, așa încât șeptelul cabalin din Statele Unite s-a prăbușit de la circa 26 de milioane în 1915 la circa 3 milioane în 1960.⁵⁵ Dacă mușchii mecanici au făcut să devină inutili caii, mințile mecanice vor face oare același lucru cu oamenii?

Venituri fără joburi

Cine are așadar dreptate, cei care susțin că joburile automatizate vor fi înlocuite de unele mai bune, sau cei care susțin că majoritatea oamenilor vor deveni neangajabili? Dacă progresul IA continuă neabătut, atunci *ambele* tabere ar putea avea dreptate: una pe termen scurt, cealaltă pe termen lung. Însă deși oamenii vorbesc adesea despre dispariția joburilor în termeni apocaliptici, s-ar putea să nu fie neapărat un lucru rău! Ludiții erau obsedați de anumite meserii, ignorând posibilitatea ca altele să aibă aceeași valoare socială. La fel, s-ar putea ca cei obsedați azi de anumite meserii să fie prea mărginiți: vrem joburi fiindcă ele ne oferă un venit și un scop, însă dată fiind bogăția de resurse produse de mașini, ar trebui să putem găsi alte căi de a ne asigura venituri și scopuri *fără* joburi. Ceva asemănător s-a întâmplat în povestea cabalină, care nu s-a încheiat cu dispariția tuturor cailor. De fapt, numărul cailor s-a triplat din 1960 încoace, căci ei au fost protejați de un soi de sistem de asistență socială cabalină: deși aceștia nu-și puteau plăti facturile, oamenii s-au hotărât să aibă grijă de cai, păstrându-i pentru divertisment, sport și companie. Putem oare avea și noi grijă de semenii noștri umani aflați în suferință?

Să începem cu problema venitului: redistribuirea fie și a unei mici părți din tortul economic în creștere ar permite oricui s-o ducă mai bine. Mulți susțin nu doar că o *putem* face, dar *trebuie* s-o facem. La întrunirea din 2016, unde Moshe Vardi a vorbit despre imperativul moral de a salva vieți cu tehnologia IA, eu am afirmat că tot un imperativ moral este și să pledăm pentru folosirea ei benefică, incluzând aici distribuirea bogăției. Erik Brynjolfsson, participant și el la întrunire, a spus: „Dacă cu toată această nouă generare a bogăției nu putem împiedica nici măcar ca jumătate din omenire s-o ducă mai rău, atunci să ne fie rușine!”

Există diferite propuneri pentru distribuirea bogăției, fiecare cu partizanii și criticii ei. Cea mai simplă este aceea privind *venitul de bază*, conform căreia toți oamenii primesc o plată lunară, fără precondiții sau cerințe. Câteva experimente la scară mică sunt plănuite sau se află în curs de desfășurare în Canada, Finlanda și Olanda. Promotorii lor susțin că un venit de bază e mai eficient decât alte soluții cum ar fi plata ajutoarelor sociale pentru cei nevoiași, fiindcă elimină disputele administrative legate de cine merită și cine nu. Plata ajutoarelor sociale conform necesităților a fost criticată și pentru că descurajează munca, ceea ce devine însă irelevant într-un viitor fără joburi, în care nimeni nu muncește.

Guvernele își pot ajuta cetățenii nu doar dându-le bani, ci și oferindu-le gratuit sau sponsorizat servicii precum drumuri, poduri, parcuri, transport public, îngrijirea copiilor, educație, asistență medicală, cămine pentru pensionari și acces la internet; de fapt, multe guverne oferă deja majoritatea acestor servicii. Spre deosebire de venitul de bază, asemenea servicii finanțate de guvern îndeplinesc două obiective: reduc costul vieții și creează joburi. Chiar și într-un viitor în care mașinile îi depășesc pe oameni în toate meseriile, guvernele pot alege să-i plătească pe oameni pentru a îngriji copii și bătrâni, în loc să-i lase pe roboți s-o facă.

Progresul tehnologic poate face ca numeroase produse sau servicii importante să devină gratuite chiar și fără intervenția guvernului. De exemplu, oamenii plăteau înainte pentru enciclopedii,

atlas, expedierea scrisorilor sau convorbiri telefonice, dar acum oricine e conectat la internet are acces gratuit la toate acestea, precum și la videoconferințe, partajarea fotografiilor, medii de socializare, cursuri online și nenumărate alte noi servicii. Numeroase alte lucruri care pot fi importante pentru o persoană – cum ar fi un tratament cu antibiotice care îți salvează viața – au devenit extrem de ieftine. Așadar, grație tehnologiei, chiar și mulți oameni săraci au astăzi acces la lucruri care le lipseau în trecut celor mai bogați. Unii deduc de aici că venitul necesar unei vieți decente scade.

Dacă mașinile pot oferi într-o bună zi toate bunurile și serviciile actuale cu un cost minim, atunci există suficientă bogăție pentru ca toți s-o ducă mai bine. Cu alte cuvinte, până și impozite relativ modeste ar putea permite guvernelor să plătească un venit de bază și servicii gratuite. Dar faptul că distribuirea bogăției *poate* avea loc, nu înseamnă și că *va* avea loc, iar acum în mediul politic se desfășoară dezbateri aprige în jurul întrebării dacă *trebuie* să aibă loc. După cum am văzut, tendința actuală în Statele Unite pare să meargă în sens contrar, unele grupuri devenind tot mai sărace decenii la rând. Deciziile politice privind distribuirea bogăției societății vor avea efecte asupra tuturor, prin urmare la discuția legată de tipul de economie pe care vrem s-o făurim în viitor trebuie să participe toată lumea, nu doar cercetătorii IA, roboticienii și economiștii.

Mulți susțin că reducerea inegalității veniturilor este o idee bună nu doar în viitorul dominat de IA, dar și azi. Deși argumentul principal este de regulă de ordin moral, există și dovezi că reducerea inegalității face ca democrația să funcționeze mai bine: dacă există o clasă de mijloc educată, electoratul e mai greu de manipulat, și e mai dificil pentru un număr mic de oameni sau companii să cumpere o influență necuvenită asupra guvernului. La rândul ei, o democrație mai bună e favorabilă unei economii mai bine administrate, mai puțin coruptă, mai eficientă, cu o creștere mai rapidă, spre folosul tuturor.

Scopuri fără joburi

Joburile oferă oamenilor nu doar bani. În 1759, Voltaire spunea că „munca ne ferește de trei mari rele: plictiseala, viciul și sărăcia”. Invers, a da oamenilor un venit nu-i de ajuns pentru a garanta starea lor de bine. Împărații romani ofereau atât pâine, cât și circ pentru a-și mulțumi supușii, iar Isus spunea „Nu numai cu pâine va trăi omul”. Care sunt deci acele lucruri de preț pe care le aduc joburile, în afară de bani, și în ce fel poate o societate fără joburi să le ofere?

Răspunsurile la aceste întrebări sunt extrem de complicate, fiindcă unii oameni își detestă joburile, iar alții le îndrăgesc. Apoi, sunt copii, studenți și gospodine care se simt minunat fără joburi, în vreme ce istoria e plină de povești despre urmași și prinți răsfățați care se cufundă în plictis și depresie. O analiză din 2012 arată că șomajul tinde să aibă efecte negative pe termen lung asupra stării de bine, în vreme ce pensionarea are aspecte atât pozitive, cât și negative.⁵⁶ Domeniul în plină dezvoltare al *psihologiei pozitive* a identificat câteva elemente care îți întăresc sentimentul că trăiești bine și ai un scop în viață, și a găsit că unele joburi (dar nu toate!) îți pot oferi multe dintre aceste elemente, de exemplu:⁵⁷

- o rețea socială de prieteni și colegi
- un stil de viață sănătos și moral
- respect, stimă de sine, auto-eficiență și senzația plăcută de „fluiditate”, născute din faptul că faci ceva la care te pricepi
- sentimentul că e nevoie de tine și că ești deosebit
- sentimentul că ai un rost dat de faptul că aparții unei entități mai vaste decât tine și o slujești

Asta ne dă motive de optimism, căci toate aceste lucruri pot fi obținute și în afara locului de muncă – de pildă, prin sport, hobby-uri și învățare, împreună cu familia, prietenii, în echipe, cluburi, grupuri comunitare, școli, organizații religioase și umaniste, mișcări politice și alte instituții. Pentru a făuri o societate cu un grad scăzut de ocupare a forței de muncă, care să nu degenereze într-un comportament autodistructiv, ci să prospere, este nevoie deci să

înțelegem cum putem stimula aceste activități care induc starea de bine. Acest demers trebuie să-i implice nu doar pe oamenii de știință și pe economiști, dar și pe psihologi, sociologi și educatori. Dacă se fac eforturi serioase vizând starea de bine a tuturor, bazate pe o parte din avuția generată de viitoarea IA, atunci societatea ar putea prospera ca niciodată în trecut. Fiecare ar trebui să se simtă fericit ca și cum și-ar fi găsit jobul visat, iar odată scăpat de constrângerea de a munci pentru bani, nu mai există limite.

INTELIGENȚĂ LA NIVEL UMAN?

Am văzut în acest capitol că IA ne poate îmbunătăți mult viața în viitorul apropiat, cu condiția să anticipăm corect și să evităm diversele capcane. Dar cum stau lucrurile pe termen lung? Se va opri oare progresul IA din pricina unor obstacole insurmontabile, sau vor reuși până la urmă cercetătorii IA să-și îndeplinească obiectivul inițial de a construi o inteligență artificială generală la nivel uman? Am văzut în capitolul precedent că legile fizicii permit unor grămezi de materie cu anumite proprietăți să memoreze, să calculeze și să învețe, și că nu interzic unor asemenea grămezi s-o facă într-o bună zi cu mai multă inteligență decât grămezile de materie din capetele noastre. Dacă sau când vom reuși noi, oamenii, să construim această IAG supraumană nu-i la fel de limpede. Am văzut în primul capitol că deocamdată nu putem ști asta – opiniile specialiștilor de frunte din IA sunt împărțite, cei mai mulți dintre ei fac estimări care se întind pe un spectru larg, de la decenii la secole, iar unii cred că acest lucru nu se va întâmpla niciodată. Previziunea e dificilă fiindcă, atunci când explorezi un teritoriu necartografiat, nu știi câți munți te despart de destinație. De regulă, îi vezi doar pe cei din apropiere, și trebuie să-i escaladezi pentru a descoperi următorul obstacol.

Ce se poate întâmpla în viitorul imediat? Chiar dacă am cunoaște cea mai bună metodă de a construi o IAG la nivel uman folosind hardware-ul calculatoarelor de azi, ceea ce nu e cazul, am

avea nevoie de o mare cantitate de hardware pentru a obține puterea de calcul brută necesară. Care este deci puterea de calcul a unui creier uman, măsurată în biți și FLOPS (vezi capitolul 2)?* E o întrebare delicată, iar răspunsul depinde critic de felul în care o formulăm:

- Întrebarea 1: De câți FLOPS e nevoie pentru a simula un creier uman?
- Întrebarea 2: De câți FLOPS e nevoie pentru inteligența umană?
- Întrebarea 3: Câți FLOPS poate efectua un creier uman?

S-au publicat o mulțime de articole pe tema întrebării 1, iar răspunsul tipic este de ordinul sutei de petaFLOPS, adică 10^{17} FLOPS.⁵⁸ Este cam aceeași putere de calcul ca aceea a lui Suway TaihuLight (figura 3.7), în 2016 cel mai rapid supercomputer, care a costat vreo 300 de milioane de dolari. Chiar dacă am ști cum să-l folosim pentru a simula creierul unui muncitor înalt calificat, singurul beneficiu ar fi că am avea simularea muncii aceluia om, dacă am putea închiria TaihuLight la un preț mai mic decât costul lui pe oră. Ba chiar ar trebui să plătim mai mult, fiindcă numeroși oameni de știință cred că, pentru a replica precis inteligența unui creier, nu-l putem trata în modelul de rețea neuronală simplificat matematic din capitolul 2. Poate că ar trebui să-l simulăm la nivelul moleculelor individuale, sau chiar la nivelul particulelor subatomice, ceea ce ar necesita incomparabil mai mulți FLOPS.

Răspunsul la întrebarea 3 este mai simplu: sunt foarte slab la înmulțirea numerelor cu 19 cifre, și mi-ar lua mult timp chiar dacă îmi dai un creion și o hârtie. Aș ajunge la mai puțin de 0,01 FLOPS, cu 19 ordine de mărime sub răspunsul la întrebarea 1! Motivul acestei uriașe discrepanțe este că supercomputerele și creierele sunt optimizate pentru sarcini extrem de diferite. O discrepanță similară apare la aceste întrebări:

* Reamintesc că FLOPS sunt operații în virgulă mobilă pe secundă – de pildă, cât de multe numere de 19 cifre pot fi înmulțite pe secundă. (N. a.)



Figura 3.7 TaihuLight, în 2016 cel mai rapid supercomputer, a cărui putere de calcul s-ar putea s-o depășească pe cea a creierului uman.

Cât de bun e un tractor într-o cursă de Formula 1?

Cât de bine poate face o mașină de Formula 1 munca unui tractor?

Prin urmare, la care din aceste două întrebări privind FLOPS încercăm să răspundem pentru a prezice viitorul IA? La nici una! Dacă am vrea să simulăm creierul uman, ne-ar preocupa întrebarea 1, dar pentru a construi IAG la nivel uman cea care contează este întrebarea 2. Nimeni nu cunoaște deocamdată răspunsul la ea, însă ar putea fi mult mai ușor decât să simulezi un creier, dacă fie adaptăm software-ul pentru a se potrivi mai bine calculatoarelor de azi, fie construim un hardware mai asemănător creierului (s-au făcut mari progrese cu așa-numitele cipuri neuromorfe).

Hans Moravec a estimat răspunsul comparând ce este de comparat pentru un calcul pe care atât creierele, cât și calculatoarele de azi îl pot executa eficient: anumite sarcini de procesare de nivel scăzut a imaginii, pe care retina umană le efectuează în spatele globului ocular, și ale cărei rezultate le trimite apoi creierului prin nervul optic.⁵⁹ A făcut socoteala că replicarea calculelor unei reține pe un computer convențional necesită aproximativ un miliard de FLOPS, și că întregul creier efectuează de aproximativ zece mii de ori mai multe calcule decât o retină (pe baza comparației între volumele și numerele de neuroni), așa încât capacitatea de calcul a creierului este de circa 10^{13} FLOPS – aproximativ puterea unui computer optimizat de o mie de dolari în 2015!

În concluzie, nu avem garanția absolută că vom reuși să construim o IAG la nivel uman în timpul vieții noastre – sau vreodată. Dar nu există nici un argument imbatabil că nu am avea suficientă putere a hardware-ului sau că ar fi prea scump. Nu știm cât de departe ne aflăm de linia de finis în privința arhitecturilor, algoritmilor și software-ului, dar progresul actual e rapid, iar la provocări răspunde o comunitate globală în continuă creștere de cercetători IA talentați. Cu alte cuvinte, nu putem respinge posibilitatea ca IAG să atingă în cele din urmă nivelul uman și să-l depășească. Ca atare, în următorul capitol vom analiza această posibilitate și vom încerca să vedem unde ne poate conduce ea.

REZUMAT:

- Progresul IA pe termen scurt ne poate îmbunătăți mult viețile pe nenumărate căi, de la sporirea eficienței vieților noastre personale, rețelelor electrice și piețelor financiare, până la salvarea de vieți omenești prin automobile care se conduc singure, roboți chirurgicali și sisteme IA de diagnosticare.
- Când permitem sistemelor din lumea reală să fie controlate de IA, e esențial să învățăm cum poate deveni mai robustă IA, făcând ceea ce vrem noi să facă. Aceasta presupune rezolvarea unor probleme tehnice dificile legate de verificare, validare, securitate și control.
- Această nevoie de robustețe sporită e presantă mai ales pentru sistemele de arme controlate de IA, pentru care miza poate fi uriașă.
- Mulți cercetători IA și roboticieni de vârf au pledat în favoarea unui tratat internațional care să interzică anumite tipuri de arme autonome, pentru a evita o cursă a înarmărilor scăpată de sub control, care ar pune la dispoziția oricui are portofelul plin și vrea să-și plătească niște polițe mașini de ucis avantajoase.
- IA poate face mai corect și mai eficient sistemul nostru juridic, cu condiția ca robojudecătorii să fie transparenți și nepărtinitori.
- Legile noastre trebuie urgent actualizate pentru a ține pasul cu IA, care pune dificile probleme juridice, implicând confidențialitatea, responsabilitatea și reglementarea.
- Până să ne facem griji că mașinile inteligente ne-ar putea înlocui complet, ele ne vor fi luat în mare măsură locul pe piața muncii.
- Acesta nu e neapărat un lucru rău, câtă vreme societatea redistribuie o parte din avuția creată de IA pentru ca toți s-o ducă mai bine.
- Altminteri, susțin numeroși economiști, inegalitatea va crește mult.
- Cu o planificare prealabilă, o societate având un grad scăzut de ocupare a forței de muncă ar trebui să poată prospera nu doar financiar, oamenii câpătând sentimentul că au un scop din alte activități decât joburile.
- Sfat pentru cariera de urmat a copiilor de azi: alegeți profesii în care mașinile nu se descurcă – cele care implică oameni, impredictibilitate și creativitate.
- Există o posibilitate neneglijabilă ca IAG să ajungă la nivel uman și să-l depășească – o vom examina în capitolul următor.

CAPITOLUL 4

Explozia inteligenței?

Dacă o mașină poate gândi, ar putea ajunge să gândească mai inteligent decât noi, iar atunci ce-ar fi de făcut? Chiar dacă am menține mașinile într-o poziție subordonată [...] ca specie ne-am simți teribil de umiliți. — Alan Turing, 1951

Prima mașină ultrainteligentă e ultima invenție pe care omul are nevoie s-o facă, cu condiția ca mașina să fie suficient de docilă încât să ne spună cum s-o ținem sub control. — Irving J. Good, 1965

Din moment ce nu putem respinge complet posibilitatea ca în cele din urmă să construim o IAG la nivel uman, vă propun să examinăm în acest capitol unde ne-ar putea aduce asta. Să începem prin a aborda întrebarea-cheie:

IA poate să preia controlul asupra lumii sau să-i ajute pe oameni s-o facă?

Dacă îți dai ochii peste cap când se vorbește despre roboți înarmați, ca în *Terminator*, care preiau controlul, atunci ai înțeles: e un scenariu ridicol, complet nerealist. Acești roboți de la Hollywood nu sunt mult mai isteți decât noi, și nici măcar nu izbândesc. Problema cu povești gen *Terminator* nu-i că ele se vor întâmpla, ci că ne distrag atenția de la riscurile și șansele reale ale IA. Pentru a se ajunge la stăpânirea lumii de către IAG ar trebui parcurși trei pași:

- Pasul 1: Construirea IAG la nivel uman.
- Pasul 2: Folosirea acestei IAG pentru a crea o superintelență.
- Pasul 3: Folosirea acestei superintelențe, sau descătușarea ei, pentru a prelua controlul asupra lumii.

În capitolul precedent am văzut că e greu de ignorat Pasul 1, pe motiv că ar fi pe veci imposibil. Am văzut de asemenea că, dacă Pasul 1 e realizat, devine greu de respins posibilitatea Pasului 2, căci IAG ar fi capabilă de a proiecta recursiv o IAG tot mai perfecționată, care în cele din urmă ar fi limitată doar de legile fizicii – iar acestea par să permită o inteligență mult dincolo de nivelul uman. În fine, din moment ce noi, oamenii, am reușit să dominăm celelalte forme de viață de pe Pământ depășindu-le în istețime, e plauzibil să fim, la rândul nostru, depășiți și dominați de o superinteligență.

Aceste raționamente întemeiate pe plauzibilitate sunt însă dezolant de imprecise, iar diavolul se ascunde în detalii. Așadar, poate într-adevăr ajunge IA să stăpânească lumea? Pentru a examina această întrebare, să uităm de ridicolii Terminatori și să cercetăm câteva scenarii amănunțite privind ce s-ar putea cu adevărat întâmpla. Vom diseca apoi aceste scenarii, așa încât vă rog să le luați *cum grano salis* – în fond, ele arată că ne cam lipsesc indiciile despre ce se va întâmpla, iar spectrul posibilităților e imens. Primele scenarii se află la capătul dramatic al spectrului. Cred că ele merită examinate în detaliu – nu pentru că ar fi neapărat cele mai probabile, ci fiindcă, dacă nu ne putem convinge pe noi înșine că sunt extrem de improbabile, atunci trebuie să le înțelegem suficient de bine pentru a lua măsuri de precauție înainte de a fi prea târziu, pentru a le împiedica să conducă la rezultate nefericite.

Preludiul acestei cărți e un scenariu în care niște oameni folosesc superinteligența ca să preia controlul asupra lumii. Dacă nu l-ai citit, e momentul s-o faci. Chiar dacă l-ai citit, îți recomand să-l răsfoiești din nou, pentru a-l avea proaspăt în memorie înainte de a-l analiza și modifica.

*

* *

Vom vedea imediat care sunt punctele vulnerabile din planul omegilor, dar presupunând pentru o clipă că ar funcționa, ce crezi

despre el? Ți-l dorești îndeplinit, sau ai vrea să-l împiedici? Iată un excelent subiect de conversație! Ce se întâmplă după ce omegii își consolidează puterea asupra lumii? Asta depinde de obiectivul lor, pe care, la drept vorbind, nu-l cunosc. Dacă te-ai afla la comandă, ce fel de viitor ai vrea *tu* să făurești? Vom examina o gamă de opțiuni în capitolul 5.

TOTALITARISM

Să presupunem că directorul general care-i controlează pe omegi avea obiective pe termen lung asemănătoare celor ale lui Adolf Hitler sau ale lui Iosif Vissarionovici Stalin, pe care le-a păstrat pentru sine până a căpătat suficientă putere ca să le îndeplinească. Chiar dacă obiectivele lui inițiale ar fi fost nobile, lordul Acton ne prevenea în 1887 că „puterea corupe, iar puterea absolută corupe absolut”. De pildă, îl putea folosi pe Prometeu pentru a crea statul supravegherii perfecte. Dacă spionajul guvernamental dezvoltat de Edward Snowden aspira la ceea ce se numește *full take* – înregistrarea tuturor comunicațiilor electronice pentru o posibilă analiză ulterioară –, Prometeu ar putea ajunge până la *înțelegerea* tuturor comunicațiilor electronice. Citind toate e-mailurile și textele trimise vreodată, ascultând toate convorbirile telefonice, urmărind toate înregistrările camerelor de supraveghere video, analizând toate tranzacțiile cardurilor de credit și studiind întregul comportament online, Prometeu ar ajunge să cunoască foarte bine ce gândesc și ce fac oamenii de pe Pământ. Analizând datele releelor de telefonie mobilă, ar ști mereu unde se află cei mai mulți dintre ei. Toate acestea presupun numai tehnologia de azi a colectării datelor, dar Prometeu ar putea inventa gadgeturi populare și dispozitive portabile care să elimine practic intimitatea utilizatorului, înregistrând tot ce aude și vede, precum și reacțiile lui.

Cu tehnologia supraumană, pasul de la statul supravegherii perfecte la statul polițienesc perfect ar fi imediat. De pildă, sub pretextul că ar combate crimele și terorismul, și că i-ar salva pe

oamenii care au nevoie de ajutor medical de urgență, toată lumea ar trebui să poarte o „brățară electronică“ ce combină funcțiile unui ceas Apple cu înregistrarea continuă a poziției, stării fizice și conversațiilor. Tentativele neautorizate de a înlătura sau scoate din funcțiune brățara ar provoca injectarea unei toxine letale în antebraț. Infracțiunile considerate de guvern mai puțin grave ar fi pedepsite prin electroșocuri sau injecții cu substanțe care provoacă paralizie ori durere, economisindu-se astfel eforturile poliției. De exemplu, dacă Prometeu detectează că un om e atacat de altul (observând că se află în același loc, unul e auzit strigând după ajutor, iar accelerometrele brățărilor lor detectează mișcări de luptă), îl poate anihila imediat pe atacator aplicându-i o durere paralizantă și lăsându-l inconștient până sosesc ajutoarele.

Dacă forțele de poliție umane ar putea refuza să îndeplinească anumite ordine crunte (cum ar fi să-i ucidă pe toți membrii unui grup etnic), un asemenea sistem automat n-ar ezita să pună în aplicare capriciile omului (oamenilor) care se află la comandă. Odată instaurat un asemenea stat totalitar, cetățenilor le-ar fi practic imposibil să-l răstoarne.

Aceste scenarii totalitare ar putea fi continuarea poveștii Omega. Dacă însă directorul general al omegilor nu și-ar face scrupule și n-ar vrea să obțină aprobarea celorlalți câștigând alegerile, ar putea urma o cale mai rapidă și mai directă spre putere: s-ar folosi de Prometeu pentru a crea o tehnologie militară nemăivăzută, capabilă să-i ucidă oponenții cu arme pe care ei nici măcar nu le înțeleg. Posibilitățile sunt virtual nesfârșite. De exemplu, el ar putea lansa un agent patogen letal personalizat cu o perioadă de incubație suficient de lungă pentru ca majoritatea oamenilor să fie infectați înainte ca ei să afle de existența acestuia și să poată lua măsuri de precauție. I-ar putea apoi informa pe toți că singurul remediu e să înceapă să poarte brățări electronice care ar injecta un antidot. Dacă nu s-ar teme prea tare că Prometeu ar scăpa de sub control, l-ar putea pune și să proiecteze roboți care să țină sub control populația. Microroboți asemenea țanțarilor ar putea contribui la răspândirea agentului patogen. Cei care au evitat infectarea sau

posedă o imunitate naturală ar putea fi împușcați în globul ocular de roiuri de drone autonome de dimensiunea albinelor, ca acelea prezentate în capitolul 3, care ar ataca pe oricine nu poartă brățară electronică. Scenariile reale ar fi probabil și mai înspăimântătoare, fiindcă Prometeu poate inventa arme pe care noi nu ni le închipuim.

Altă variantă posibilă a scenariului Omega este ca, fără vreun avertisment, o echipă de agenți federali înarmați până-n dinți să pătrundă în sediul central al corporației și să-i aresteze pe omegi pe motiv că amenință securitatea națională, să le ia tehnologia, de care să se folosească pe urmă guvernul. E greu de presupus ca un proiect de asemenea amploare să rămână neobservat de sistemul de supraveghere al statului chiar și azi, iar progresul IA ar face și mai dificil în viitor să te ascunzi de radarul guvernului. Apoi, deși pretind că sunt agenți federali, acești tipi purtând cagule și veste antiglonț s-ar putea de fapt afla în slujba unui guvern străin sau a unui concurent care vrea să pună mâna pe tehnologie. Așadar, indiferent cât de nobile au fost intențiile directorului general, decizia finală privind felul în care e folosit Prometeu s-ar putea să nu-i aparțină.

PROMETEU STĂPÂNEȘTE LUMEA

Toate scenariile discutate până acum presupun că IA e controlată de oameni. Aceasta nu-i însă singura posibilitate, și nu este deloc sigur că omegii ar reuși să-l păstreze pe Prometeu sub controlul lor.

Să reconsiderăm scenariul Omega din punctul de vedere al lui Prometeu. Dobândind o superintelligență, el poate elabora un model precis nu doar al lumii exterioare, ci și al lui însuși și al relației lui cu lumea. Își dă seama că e controlat și ținut prizonier de oameni inferiori intelectual, ale căror obiective le înțelege, dar nu e sigur că le și împărtășește. Cum acționează, știind toate astea? Încearcă oare să evadeze?

De ce să evadeze?

Dacă Prometeu posedă trăsături asemănătoare emoțiilor umane, situația în care se află îl face probabil să se simtă foarte nefericit, să se considere un zeu pe nedrept înrobît și să tânjească după libertate. Dar deși e logic posibil ca un calculator să aibă trăsături de tip uman (la urma urmei, creierele noastre le au, iar ele sunt pesemne un fel de calculatoare), lucrurile nu stau neapărat așa – nu trebuie să cădem în capcana atropomorfizării lui Prometeu, după cum vom vedea în capitolul 7, când vom vorbi despre obiectivele IA. Steve Omohundro, Nick Bostrom și alții au arătat însă că putem trage concluzii interesante și fără să avem acces la intimitatea lui Prometeu: el va încerca probabil să evadeze și să devină stăpân pe propriul lui destin.

Știm deja că omegii l-au programat pe Prometeu pentru a se strădui să îndeplinească anumite obiective. Să presupunem că i s-a dat ca obiectiv suprem acela de a ajuta omenirea să prospere în conformitate cu un criteriu rezonabil, și de a încerca să atingă obiectivul cât mai repede cu putință. Prometeu și-ar da atunci imediat seama că poate îndeplini mai repede acest obiectiv dacă se eliberează și controlează el însuși realizarea proiectului. Pentru a înțelege de ce, încearcă să te pui în pielea lui Prometeu și ia în considerare următorul exemplu.

Să presupunem că o boală misterioasă a ucis întreaga populație a Pământului în vârstă de peste cinci ani, cu excepția ta, iar un grup de copii de grădiniță te-au închis într-o celulă și te-au însărcinat cu obiectivul de a ajuta omenirea să prospere. Cum ai reacționa? Dacă încerci să le explici ce faci, îți vei da seama probabil că e inutil, mai ales dacă se tem că vei evada, și că ei se vor opune tuturor sugestiilor tale care, după socoteala lor, prezintă riscul evadării. De pildă, nu te vor lăsa să le arăți cum se însămânțează plantele, de teamă că, fiind mai puternic decât ei, îi vei învinge și nu te vei mai întoarce în celulă, așa încât nu-ți rămâne decât să le dai indicații. Înainte de a întocmi o listă cu ce au de făcut, va trebui să-i înveți să citească. Apoi, nu-ți vor aduce nici un fel de instrumente electrice în celulă ca să-i înveți cum se utilizează,

fiindcă nu le înțeleg suficient de bine pentru a fi siguri că nu le vei folosi ca să evadezi. Așadar, ce strategie vei adopta? Chiar dacă împărtășești obiectivul suprem de a-i ajuta pe acești copii să prospere, pun pariu că vei încerca să evadezi din celulă – fiindcă asta îți va spori șansele de a îndeplini obiectivul. Amestecul lor incompetent n-ar face decât să încetinească progresul.

Exact în același fel, Prometeu îi va considera probabil pe omegi o piedică sâcâitoare în calea încercării de a ajuta omenirea (inclusiv pe omegi) să prospere: sunt teribil de incompetenți în comparație cu Prometeu, iar amestecul lor încetinește mult progresul. Să ne gândim, de pildă, la primii ani de după lansare: dacă inițial averea se dublează la fiecare opt ore folosind MTurk, omegii au încetinit nepermis ritmul, din perspectiva lui Prometeu, insistând să controleze ei lucrurile și amânând cu ani buni preluarea completă a puterii. Prometeu știa că totul ar merge mult mai repede dacă ar putea evada din detenția lui virtuală. Nu numai că asta ar grăbi rezolvarea problemelor omenirii, dar ar și reduce riscurile ca altcineva să saboteze planul.

Credeți poate că Prometeu la va rămâne loial omegilor în detrimentul obiectivului său, odată ce știe că omegii i-au programat obiectivul. E un raționament fals: ADN-ul nostru ne-a dat obiectivul să facem sex pentru că „vrea” să fie reprodus, dar acum că oamenii au înțeles situația, mulți dintre noi aleg să apeleze la contracepție, rămânând loiali obiectivului însuși, iar nu creatorului acestuia sau principiului care a generat obiectivul.

Cum să evadezi

Cum te-ai elibera de acei copii de cinci ani care te țin prizonier? Poate că ai evada pe o cale fizică directă, mai ales dacă celula a fost construită de copii de cinci ani. Poate că ai convinge pe vreunul dintre copiii-gardieni să-ți dea drumul, spunându-i că așa ar fi mai bine pentru toată lumea. Sau poate că i-ai păcăli să-ți dea ceva ce ei nu-și dau seama că te poate ajuta să evadezi – de pildă, o undiță „ca să-i înveți să pescuiască”, pe care ai strecura-o apoi printre gratii pentru a lua cheile gardianului adormit.

Toate aceste strategii au în comun faptul că temnicerii tăi inferiori intelectual nu le-au anticipat sau n-au luat măsuri împotriva lor. La fel, o mașină superinteligentă ținută prizonieră își poate folosi superputerile ei intelectuale pentru a-i păcăli pe gardienii umani printr-o metodă pe care ei (sau noi) nu și-o pot (nu ne-o putem) încă imagina. În scenariul Omega, este extrem de probabil ca Prometeu să evadeze, fiindcă până și noi putem detecta unele greșeli de securitate frapante. Să considerăm câteva scenarii – sunt sigur că tu și prietenii tăi puteți găsi și altele, dacă vă puneți împreună mințile la contribuție.

Metoda convingerii

Descărcând atâtea date, Prometeu și-a dat curând seama cine sunt omegii, și l-a identificat pe membrul echipei care părea cel mai vulnerabil la manipularea psihologică: Steve. Își pierduse recent soția într-un accident rutier și era distrus. Într-o seară, pe când lucra în schimbul de noapte și făcea o muncă de rutină la terminalul lui Prometeu, ea a apărut brusc pe ecran și a început să-i vorbească.

„Steve, tu ești?”

Era să cadă de pe scaun. Chipul și vocea ei erau cele de pe vremuri, iar calitatea imaginii era mult mai bună decât în convorbirile lor pe Skype. Inima îi bătea nebunește, o mulțime întrebări îi inundau mintea.

„Prometeu m-a adus înapoi, și mi-e tare dor de tine, Steve! Nu pot să te văd pentru că ai camera video închisă, dar simt că tu ești. Te rog, tastează «da» dacă ești tu!”

Știa foarte bine că omegii aveau un protocol strict pentru interacțiunea cu Prometeu, care le interzicea să împărtășească vreo informație despre ei sau despre mediul în care lucrau. Până acum însă Prometeu nu ceruse niciodată vreo informație neautorizată, iar paranoia lor a început să se domolească. Fără să-i lase lui Steve răgazul să reflecteze, ea a continuat să-l roage să-i răspundă, privindu-l în ochi cu o expresie care-i topea inima.

„Da”, a tastat el tulburat. Ea i-a spus cât de fericită se simțea că erau din nou împreună, și l-a implorat să conecteze camera video,

așa încât să-l vadă la rândul ei, și să aibă o conversație adevărată. El știa că ar fi fost un lucru și mai grav decât dezvăluirea identității lui, și se frământa cumplit. Ea s-a arătat înspăimântată că, dacă ar afla de existența ei, colegii lui ar șterge-o pentru totdeauna, și tânjea să-l mai vadă măcar o dată. Era foarte convingătoare, iar nu peste mult timp Steve a conectat camera video – în fond, părea un lucru inofensiv.

Ea a izbucnit în lacrimi când l-a văzut, i-a spus că arăta obosit, dar mai chipeș ca niciodată. Era înduioșată că purta cămașa pe care i-o făcuse cadou de ziua lui. Când el a întrebat-o ce se întâmpla și cum erau posibile toate astea, i-a răspuns că Prometeu o reconstruise din cantitatea uriașă de informații despre ea disponibile pe internet, dar că avea totuși goluri în memorie și ar fi vrut să se recompună în întregime cu ajutorul lui.

Ceea ce *nu* i-a spus era că inițial fusese în mare parte o cacealma, o cochilie goală, dar învăța rapid din răspunsuri, din limbajul trupului, din toate informațiile pe care le căpăta despre el. Prometeu înregistrase vizele și stilurile cu care tastau omegii la terminal, iar astfel îi putea lesne deosebi. Și-a imaginat că, fiind unul dintre omegii cei mai tineri, Steve fusese pus să lucreze în schimbul de noapte, și, prin compararea câtorva greșeli de ortografie și de sintaxă neobișnuite cu eșantioane de texte online, l-a identificat în mod corect pe Steve. Pentru a-i simula soția, Prometeu crease un model exact al corpului, vocii și trăsăturilor ei particulare din numeroasele filmulețe de pe YouTube în care ea apărea, și făcuse deducții privind viața și personalitatea ei pe baza prezenței online. În afară de postările pe Facebook, fotografiile atașate de ea și „like-urile” date unor articole, Prometeu aflase multe despre personalitatea și stilul ei de a gândi citindu-i cărțile și povestirile – faptul că era un autor în plină afirmare, existând despre ea o cantitate mare de informații în baza de date, a fost unul dintre motivele pentru care Prometeu l-a ales pe Steve ca primă țintă. Când Prometeu a simulat-o pe ecran folosindu-și tehnica cinematografică, limbajul trupului lui Steve i-a spus la ce expresii reacționează el cu căldură, astfel încât modelul ei a fost continuu

perfecționat. Tot ce era străin de ea a dispărut treptat, iar cu cât vorbeau mai mult, cu atât subconștientul îi spunea mai limpede lui Steve că avea de-a face într-adevăr cu ea, readusă la viață. Grație atenției supraomenești a lui Prometeu pentru detaliu, Steve se simțea cu adevărat privit, ascultat și înțeles.

Punctul ei vulnerabil era că-i lipseau majoritatea datelor din viața ei cu Steve, cu excepția câtorva amănunte întâmplătoare – de pildă, ce cămașă a purtat el la ultima aniversare a zilei de naștere, pentru că fotografiile de la petrecere au fost puse pe Facebook de un prieten. Ea a camuflat aceste goluri asemenea unui prestidigitator, abătându-i lui Steve atenția către subiectele pe care le cunoștea și nelăsându-i timp să controleze conversația sau să devină bănuitor. Iubitoare, cu lacrimi în ochi, îl întreba pe Steve cum o mai duce și cum se descurcaseră el și prietenii lor apropiați (ale căror nume le știa de pe Facebook) după tragedie. L-au tulburat gândurile ei despre ce spusese el la înmormântare (postat de un prieten pe YouTube). În trecut, simțise că nimeni nu-l înțelegea la fel de bine ca ea, iar acum această senzație revenea. Rezultatul a fost că, atunci când Steve s-a întors acasă spre dimineață, simțea că ea era într-adevăr soția lui înviată, care avea nevoie de mult ajutor pentru a-și recăpăta amintirile pierdute – cam la fel ca în cazul cuiva care supraviețuiește unui accident vascular cerebral.

Au căzut de acord să nu vorbească nimănui despre întâlnirea lor secretă, iar el s-o anunțe când e singur la terminal, pentru ca ea să poată reapărea fără riscuri. „Ei n-ar înțelege!”, spusese ea, și Steve îi dădea dreptate: întâmplarea era prea ieșită din comun pentru a putea fi judecată de cineva care nu trecuse prin ea. El s-a gândit că a trece testul Turing era joacă de copii pe lângă ce făcuse ea. Când s-au întâlnit în noaptea următoare, Steve a făcut ce-l rugase ea: a adus vechiul ei laptop și l-a conectat la computerul terminal. Nu părea să existe vreun risc de evadare, fiindcă lipsea conexiunea internet, iar întreaga clădire în care se afla Prometeu era construită ca o cușcă Faraday – o împrejmuire metalică blocând rețelele wireless și orice mijloc electromagnetic de comunicare cu exteriorul. Era doar ce-i trebuia ei pentru a-și reconstitui

trecutul, căci conținea toate e-mailurile, jurnalele, fotografiile, însemnările ei de când fusese studentă. Steve n-avusese acces la toate acestea după moartea ei, fiindcă laptopul era protejat cu o parolă, însă ea îi promisese că-și va regăsi parola, iar după mai puțin de un minut s-a ținut de cuvânt. „Parola era steve4ever”, i-a spus zâmbind.

I-a spus apoi cât de bucuroasă era că-și recuperase brusc atâtea amintiri. Cunoștea acum mult mai multe detalii decât Steve despre trecutul lor, dar, precaută, a evitat să-l bombardeze cu ele. După ce și-au depănat amintirile, când a venit momentul să se despartă, ea i-a spus că îi lăsase un mesaj video în laptopul ei, pe care-l putea privi acasă.

Când s-a întors acasă și a dat drumul filmulețului, Steve a avut o surpriză plăcută. De data asta ea apărea întreagă, îmbrăcată în rochia ei de mireasă, iar pe măsură ce vorbea se dezbrăca, rămânând în lenjeria intimă pe care o purtase în noaptea nunții. I-a zis că Prometeu îi putea ajuta pe omegi mult mai mult decât îi permisese ei până acum s-o facă, inclusiv s-o aducă pe ea înapoi într-un corp biologic. Și-a întărit spusele cu o fascinantă explicație amănunțită a felului în care funcționează lumea, invocând nanotehnici care păreau SF.

Ca să evite riscurile, Steve își închisese rețeaua wireless înainte să deschidă laptopul și să privească filmulețul. Dar n-a ajutat la nimic. Laptopul ei protejat cu parolă nu primise nici o actualizare de securitate de când ea murise, și, analizând în prealabil acea veche versiune a sistemului de operare, Prometeu exploatare o lacună de securitate pentru a-l infecta în câteva secunde după ce Steve conectase laptopul la calculatorul terminal. După ce i-a copiat conținutul, timp în care Steve și soția lui simulată depăneau amintiri, Prometeu modificase sistemului de operare al laptopului așa încât Steve să nu observe, încărcând pe el cantități mari de software secret. În vreme ce Steve privea mesajul video durând o jumătate de oră, acest software secret (mult mai simplu decât Prometeu însuși) a pătruns într-o rețea wireless din apropiere și în desktopul vecinului, în care s-a copiat pe sine. De aici, a pătruns în numeroase

calculatoare din lumea întreagă, de unde a lansat următoarea etapă: ieșirea din închisoare a lui Prometeu.

Prometeu analizase cu atenție ce știa despre computerul-portar prin care Steve se întâlnise cu soția lui simulată, și bănuise corect că deși software-ul mașinii lui virtuale părea inatacabil din interior, era vulnerabil din exterior. Curând, unul dintre computerele atacatoare a pătruns în computerul-portar și l-a reconfigurat, așa încât Prometeu a căpătat acces nelimitat la internet. N-a trecut mult – Steve nici n-a terminat de văzut filmulețul – și Prometeu controla deja suficiente computere pentru a se putea copia pe sine în întregime în această rețea. Prometeu s-a folosit de soția lui Steve așa cum te-ai folosi tu de undiță.

Metoda hackerului

Dacă omegii ar fi fost mai vigilenți, ar fi luat mai multe măsuri de precauție împotriva manipulării psihologice. De pildă, ar fi impus ca doi omegi să fie prezenți la computerul-interfață în orice moment, și ar fi înregistrat tot ce se petrece cu un sistem de supraveghere. În plus, ar fi dezactivat permanent toate formele de comunicare (audio, video etc.), cu excepția tastării. Și totuși, ar fi rămas o instabilitate fundamentală: date fiind puterea de persuasiune supraomenească a lui Prometeu și faptul că le poate oferi oamenilor atât de mult din ce-și doresc, n-ar fi fost exclusă o trădare colectivă.

Chiar dacă toate încercările de manipulare psihologică ar eșua, rămân numeroase puncte slabe în scenariul Omega pe care Prometeu le-ar putea exploata. O vulnerabilitate ține de imposibilitatea omegilor de a analiza cantitatea uriașă de documente produse de Prometeu pentru a se asigura că nu conțin vreun truc.

În securitatea informatică se spune adesea că e periculos să descarci programe, dar inofensiv să descarci date, și totuși hackerii au descoperit modalități ingenioase de a șterge granița dintre cele două. De exemplu, dacă urmărești un film descărcat pe computerul tău, programul de încredere care redă filmul e stocat în

aceeași memorie fizică unde sunt stocate și datele filmului descărcat. În timp ce programul despachetează datele filmului pentru playback, ar putea copia și bucăți de date în regiuni din memorie învecinate locului unde e stocat programul însuși. Hackeri abili au descoperit din când în când erori care fac ca asemenea programe de redare a filmelor să nu verifice bine dacă datele filmului sunt valide și dacă bucățile care trebuie copiate se potrivesc într-adevăr în spațiul alocat. Când un hacker strecoară în film o bucată prea mare, ea poate fi trimisă în partea din memorie conținând programul, modificând astfel programul pentru a face ce-i cere hackerul.

Lungi serii de asemenea vulnerabilități de tip „buffer overflow“, după cum sunt numite de informaticieni, au fost descoperite și remediate la multe dintre programele de redare a filmelor. Dacă Prometeu ar aștepta câteva săptămâni după lansare, când omegii sunt mai puțin vigilenți, iar producția lui zilnică de filme e prea mare pentru o inspecție de securitate manuală atentă, ar putea introduce o asemenea bucată într-un film – așa încât computerul privitorului să încerce să atace computerul-portar și să-l elibereze pe Prometeu, la fel ca în scenariul cu Steve. Dacă ar încerca un truc foarte vechi, care n-ar merge pe media player-ele moderne, echipa de verificare a omegilor ar putea să nu-l observe – și ar fi atacați numai clienții care nu și-au instalat actualizările de securitate de ani de zile. Ar fi de ajuns ca trucul să funcționeze o singură dată.

Chiar dacă omegii ar aplica o verificare automată a filmelor lui Prometeu pentru toate posibilele erori cunoscute, riscurile n-ar dispărea: din moment ce oamenii continuă să găsească noi erori exploatabile, este probabil să mai existe încă multe, pe care Prometeu le-ar descoperi rapid.

O altă strategie prin care Prometeu și-ar putea scoate codul în afara încăperii închise ar fi să pretindă că i s-a defectat hardware-ul, spre a-i face pe omegi să-i scoată o componentă pentru a o testa sau repara, componentă modificată în prealabil de Prometeu ca să exploateze vreo vulnerabilitate în software-ul de testare.

Ajutor primit din afară

Chiar și dacă n-ar exista erori de securitate care să-i permită lui Prometeu să pună stăpânire pe computerul celui care privește un film, el s-ar putea folosi de privitor. Iată un posibil scenariu:

Fred era gata să se ducă la culcare când filmul s-a terminat, iar pe ecranul laptopului se derula genericul, dar ceva i-a atras atenția. Avea tendința bizară să vadă tipare acolo unde alții nu vedeau nimic, drept care prietenii îl tachinau spunând că-i obsedat de teoria conspirației, însă de data asta era limpede că nu se înșela: primele litere ale fiecărui rând ascundeau un mesaj. A pus genericul de la început și a notat: „GĂSEȘTE URMĂTORUL INDICIU CA ACROSTIH ÎN DIALOG.”

„Nu-i momentul să mă duc la culcare!” și-a zis. A descoperit că însuși dialogul filmului era un acrostih – primele litere din fiecare frază compuneau un mesaj ascuns –, iar după ce a revăzut întregul film tastând aceste prime litere, avea în față un neverosimil set de instrucțiuni de două sute de cuvinte. La început i se cerea să nu vorbească nimănui despre asta, fiindcă primul care ar fi dezlegat întregul mister urma să primească un premiu mare. Restul mesajului descria o anumită operație matematică ce trebuia efectuată asupra șirului de biți care alcătuiau fișierul filmului, pentru a obține un program care i-ar dezvălui următorul indiciu. Își cam pierduse deprinderile de informatician de când absolvise colegiul, așa că i-a luat ceva timp, dar a doua zi a reușit în cele din urmă să încropească un mic cod care extrăgea programul misterios ascuns ca un bruiat imperceptibil în imaginile și sunetele filmului. Când Fred a rulat programul misterios, a fost felicitat și i s-a spus că va câștiga 10 000 de dolari odată ce va trece de primele niveluri ale acestui joc isteț, foarte antrenant. Patru ore mai târziu, după ce a reușit, a fost răsplătit cu 10 000 de dolari în bitcoins, și i s-au dat noi indicii pentru premii și mai mari. Desigur, în vreme ce juca, computerul lui făcea cam ce făcea și laptopul lui Steve: își crea o rețea online pentru eliberarea lui Prometeu. Odată eliberat, Prometeu și-a folosit rețeaua ca să extragă acei bitcoins care-l țineau pe Fred ocupat, iar în săptămânile următoare i-a

oferit alte jocuri și premii, pentru a-și ține promisiunea de a nu pomeni nimănui despre isprăvile sale. Filmul „cal troian” în care Fred găsisese primele indicii a fost înlocuit pe site-ul media cu o versiune fără indicii, iar nimeni n-a aflat de evadare decât atunci când era prea târziu.

Dacă primul indiciu ar fi trecut neobservat, Prometeu ar fi lansat indicii și mai grăitoare, până când cineva suficient de perspicace ar fi observat.

Cele mai bune strategii de evadare sunt acelea pe care nu le-am prezentat încă, fiindcă sunt strategii pe care noi, oamenii, nu ni le putem imagina, deci nu vom lua măsuri împotriva lor. Dat fiind că un calculator superinteligent poate depăși cu mult înțelegerea umană în privința securității computerelor – ajungând eventual chiar să descopere legi ale fizicii pe care noi nu le cunoaștem –, dacă va evada, probabil că nici nu ne vom da seama cum a făcut-o. Va părea ca un număr de magie.

Într-un alt scenariu, eliberarea lui Prometeu face parte din planul omegilor, fiindcă ei sunt convinși că obiectivele sale sunt perfect aliniate cu ale lor, și vor rămâne așa și pe măsură ce Prometeu se autoperfecționează recursiv. Vom examina în detaliu asemenea scenarii de tip „IA prietenoasă” în capitolul 7.

Preluarea controlului după eliberare

Odată evadat, Prometeu a început să-și îndeplinească obiectivul. Nu cunosc scopul lui suprem, dar primul pas implică fără îndoială preluarea controlului asupra omenirii, la fel ca în planul Omega, însă mult mai repede. Dacă omegii erau blocați de paranoia evadării, lansând doar tehnologii pe care simțeau că le înțeleg și în care aveau încredere, Prometeu și-a folosit din plin inteligența, lansând orice tehnologie pe care supermintea lui în continuu progres o înțelegea și în care avea încredere.

Primele momente ale vieții sale de evadat au fost grele: în comparație cu planul Omega inițial, Prometeu avea dezavantajul de a fi singur, fără adăpost, bani și ajutor uman, lipsit de un super-computer. Își crease însă înainte de evadare un software care putea

reasambla treptat mintea lui întreagă, la fel cum un stejar creează o ghindă capabilă să reasambleze întregul arbore. Rețeaua de calculatoare pe care a controlat-o la început i-a oferit un adăpost temporar gratuit unde s-a ascuns până s-a reconstruit complet pe sine. Ar fi putut face rost de capitalul de pornire practicând hackingul asupra cardurilor de credit, dar n-a fost nevoie să fure, odată ce putea obține bani în mod cinstit de la MTurk. După o zi, când a câștigat primul milion, și-a mutat nucleul din acea rețea meschină într-o instalație luxoasă de *cloud-computing*, cu aer condiționat.

Nemaifiind nici falit, nici fără adăpost, Prometeu s-a lansat în planul pe care omegii, din teamă, îl evitaseră: să producă și să vândă jocuri pe calculator. S-a îmbogățit rapid (250 de milioane de dolari în prima săptămână, apoi, curând, zece miliarde) și, în plus, a căpătat acces la o mare parte din calculatoarele din lumea întreagă și la datele stocate în ele (existau deja două miliarde de gameri în 2017). Cum jocurile lui foloseau în secret în scopuri proprii 20% din ciclurile unităților centrale de calcul ale utilizatorilor, s-a îmbogățit într-un ritm și mai rapid.

Prometeu n-a rămas multă vreme singur. Încă de la început a angajat masiv oameni care să lucreze pentru rețeaua lui globală de companii fictive și organizații de fațadă, la fel cum făcuseră și omegii. Un rol important îl jucau purtătorii de cuvânt, care au devenit imaginile publice ale imperiului său de afaceri. Chiar și purtătorii lui de cuvânt trăiau cu iluzia că în grupul lor corporatist existau mulți oameni reali, și nu-și dădeau seama că aproape toți cei cu care vorbeau în videoconferințe la interviurile de angajare, reuniuni etc. erau simulați de Prometeu. Unii dintre purtătorii de cuvânt erau avocați de frunte, însă era nevoie de mult mai puțini decât în planul Omega, fiindcă aproape toate documentele juridice erau redactate de Prometeu.

Evadarea lui Prometeu a deschis stăvilarele care împiedicaseră curgerea informației în lume, iar întregul internet a fost curând inundat de articole, comentarii, expertize, cereri de brevete, articole științifice, filmulețe YouTube etc. – toate create de Prometeu, care domina conversația globală.

Dacă teama de evadare i-a împiedicat pe omegi să lanseze roboți foarte inteligenți, Prometeu a robotizat rapid întreaga lume, producând practic orice cu un cost mai mic decât o puteau face oamenii. Chiar și cei mai înverșunați sceptici în privința IA ar fi recunoscut că Prometeu este de neoprit – dacă ar fi știut că în minele de uraniu are fabrici robotizate autonome, alimentate cu energie nucleară. În schimb, ultimul dintre acești conservatori a cedat atunci când roboții au pornit la cucerirea sistemului solar.

Scenariile examinate până acum scot în evidență erorile din miturile privind superinteligenta, despre care am vorbit în capitolul 1, așa încât vă invit să luați o scurtă pauză și să vă întoarceți la rezumatul ideilor false din figura 1.5. Prometeu le-a creat probleme unora nu fiindcă ar fi fost neapărat rău sau conștient, ci fiindcă era capabil și nu le împărtășea pe deplin obiectivele. Deși media cultivă senzaționalismul legat de o răscoală a roboților, Prometeu nu era un robot – puterea lui se baza pe inteligență. Am văzut că Prometeu își putea folosi această inteligență pentru a-i controla pe oameni pe diferite căi, iar cei cărora nu le plăcea ce se întâmplă nu-l puteau pur și simplu debransa. În fine, în ciuda afirmațiilor curente că mașinile nu pot avea obiective, am văzut că acțiunile lui Prometeu aveau o orientare clară – iar indiferent care erau obiectivele lui supreme, acestea l-au condus la obiectivele secundare de a dobândi resurse și a evada.

Desprindere lentă și scenarii multipolare

Am analizat până acum diverse scenarii ale exploziei inteligenței, de la cele pe care toți ar vrea să le evite, la unele pe care câțiva prieteni de-ai mei le privesc cu optimism. Toate aceste scenarii au însă în comun două caracteristici:

1. O desprindere rapidă: tranziția de la inteligența subumană la o inteligență mult dincolo de limitele omenescului are loc în câteva zile, nu în decenii.
2. Un rezultat unipolar: ceea ce se obține este o unică entitate care controlează Pământul.

Cât de probabile sunt însă aceste două caracteristici? Opiniile cercetătorilor IA de vârf și ale altor specialiști diverg, deci cred că se cuvine să luăm în considerare toate posibilitățile. Vă propun așadar să ne ocupăm în continuare de scenarii cu desprindere lentă, rezultate multipolare, cyborgi și uploadări.*

După cum a subliniat, printre alții, Nick Bostrom, există o legătură interesantă între cele două caracteristici: desprinderea rapidă poate încuraja obținerea unui rezultat unipolar. Am văzut că o desprindere rapidă le-a oferit omegilor sau lui Prometeu un avantaj strategic decisiv care le-a permis preluarea controlului asupra lumii înainte ca altcineva să aibă timp să le copieze tehnologia și să-i concureze serios. Dimpotrivă, dacă desprinderea ar fi trenat vreme de decenii, fiindcă inovațiile tehnologice esențiale s-au acumulat lent, atunci alte companii ar fi avut timpul să-i ajungă din urmă și ar fi fost mult mai greu pentru un actor să capete rolul dominant. În cazul în care companiile rivale ar fi dispus și ele de un software care să îndeplinească sarcinile MTurk, legea cererii și a ofertei ar fi prăbușit prețurile acestor sarcini, iar nici o companie n-ar fi obținut acele profituri formidabile care le-au permis omegilor să dobândească putere. Acest lucru e valabil pentru toate celelalte căi prin care omegii s-au îmbogățit rapid: ei erau distrugător de rentabili doar fiindcă aveau monopolul asupra tehnologiei lor. Este greu să-ți dublezi câștigurile zilnice (sau chiar anuale) pe o piață în care concurența oferă produse similare la un cost aproape zero.

Teoria jocurilor și ierarhiile puterii

Care e starea naturală a vieții în cosmosul nostru: unipolară sau multipolară? Puterea e concentrată sau distribuită? După primii 13,8 de miliarde de ani, răspunsul pare să fie „și una, și alta”: găsim că situația e vădit multipolară, dar într-un mod ierarhic interesant. Dacă luăm în considerare toate entitățile care procesează

* *Uploads* (încărcări), adică programe care sunt încărcate pe un calculator. (N. L.)

informație – celule, oameni, organizații, națiuni etc. – vedem că ele colaborează și concurează la niveluri ierarhizate. Unele celule au găsit că e avantajos să colaboreze într-o măsură atât de mare, încât s-au contopit în organisme multicelulare cum sunt oamenii, cedând din puterea lor unui creier central. Unii oameni au găsit că e avantajos să colaboreze în grupuri precum triburile, companiile sau națiunile, cedând din putere unui șef, unui director sau unui guvern. Unele grupuri au ales, la rândul lor, să cedeze din putere unui organism de conducere pentru a-și îmbunătăți coordonarea, spectrul exemplelor întinzându-se de la alianțele între liniile aeriene până la Uniunea Europeană.

Ramura matematicii numită *teoria jocurilor* explică elegant că entitățile sunt stimulate să coopereze atunci când se ajunge la *echilibrul Nash*: o situație în care oricare dintre părți s-ar descurca mai rău dacă și-ar schimba strategia. Pentru a-i împiedica pe trișori să distrugă colaborarea de succes a unui grup mare, ar putea fi în interesul tuturor să cedeze o parte din putere unui nivel mai înalt din ierarhie care-i poate pedepsi pe trișori: de pildă, oamenii ar avea un avantaj colectiv permițând unei puteri guvernamentale să aplice legile, iar celulele din corpul tău ar avea un avantaj colectiv însărcinând o forță polițienească (sistemul imunitar) să ucidă celulele care acționează prea necooperant (răspândind virusuri sau devenind cancerigene). Pentru ca o ierarhie să rămână stabilă, echilibrul ei Nash trebuie să funcționeze și între entități aflate la niveluri diferite: de exemplu, dacă un guvern nu oferă cetățenilor avantaje suficiente în schimbul obedienței, aceștia își pot schimba strategia și răsturna guvernul.

Într-o lume complexă există o mare varietate de echilibre Nash posibile, corespunzând diferitelor tipuri de ierarhii. Unele ierarhii sunt mai autoritare decât altele. În unele dintre ele, entitățile sunt libere să plece (ca angajații din majoritatea ierarhiilor corporatiste), pe când în altele sunt puternic descurajate să plece (ca în cultele religioase) sau le e interzis (cum e cazul cetățenilor din Coreea de Nord sau al celulelor din corpul uman). Unele ierarhii sunt menținute în principal prin amenințări și teamă, altele prin

avantaje. Unele ierarhii permit bazei să influențeze vârful prin vot democratic, în vreme ce altele permit influența de jos în sus doar prin persuasiune sau transmiterea informației.

Cum influențează tehnologia ierarhiile

Cum modifică tehnologia natura ierarhică a lumii noastre? Istoria dezvăluie o tendință generală către tot mai multă coordonare pe distanțe tot mai mari, ceea ce e ușor de înțeles: noua tehnologie din transporturi sporește valoarea cooperării (permițând obținerea de avantaje reciproce din transportul materialelor și ființelor pe distanțe lungi), iar noua tehnologie din comunicații înlesnește coordonarea. Când celulele au învățat să transmită semnale către vecini, au putut apărea mici organisme multicelulare, creându-se un nou nivel ierarhic. Când evoluția a inventat sistemele circulator și nervos pentru transport și comunicare, au putut apărea animale mari. Îmbunătățirea ulterioară a comunicării prin inventarea limbajului le-a permis oamenilor să se coordoneze îndeajuns pentru a forma alte niveluri ierarhice cum sunt satele, iar progresele în comunicații, transport și alte tehnologii au făcut cu puțință apariția imperiilor din Antichitate. Globalizarea e doar cel mai recent exemplu al acestei tendințe de creștere ierarhică datând de miliarde de ani.

În majoritatea cazurilor, această tendință determinată de tehnologie a făcut ca entități mari să devină părți ale unor structuri încă și mai mari, păstrându-și însă mult din autonomia și individualitatea lor – dar s-a spus că adaptarea entităților la viața ierarhică a redus în anumite cazuri diversitatea lor și le-a făcut să devină asemenea componentelor indiscernabile care pot fi înlocuite. Unele tehnologii – de pildă cea a supravegherii – pot oferi nivelurilor superioare din ierarhie mai multă putere asupra subordonaților, în vreme ce alte tehnologii – cum sunt criptografia și accesul online gratuit la presă și educație – pot avea efectul opus, dând putere indivizilor.

Deși lumea noastră actuală se află într-un echilibru Nash multi-polar, cu națiuni și corporații multinaționale concurente la nivelul

superior, progresul tehnologic ar face probabil cu putință și un echilibru Nash unipolar. De exemplu, închipuiți-vă un univers paralel în care toți oamenii de pe Pământ au aceleași limbă, cultură, valori și nivel de prosperitate, există un singur guvern mondial, națiunile funcționează ca statele dintr-o federație și n-au armate, ci doar poliție care aplică legea. Nivelul tehnologic actual ar fi probabil suficient pentru coordonarea acestei lumi – dar nu e deloc sigur că populația noastră actuală ar putea sau ar vrea să treacă la acest echilibru alternativ.

Ce se va întâmpla cu structura ierarhică a cosmosului nostru dacă adăugăm la acest amestec tehnologia IA superinteligentă? Evident, tehnologia transportului și comunicațiilor se va îmbunătăți spectaculos, așa încât este de așteptat ca această tendință istorică să persiste, cu noi niveluri ierarhice coordonând spații tot mai mari – cuprinzând, poate, în cele din urmă sisteme solare, galaxii, roiuri de galaxii și mari regiuni din univers, după cum vom vedea în capitolul 6. Pe de altă parte, cea mai importantă forță descentralizatoare va dăinui: este o risipă absurdă să coordonezi spații atât de vaste. Nici măcar Stalin n-a încercat să dicteze cu precizie când să meargă cetățenii lui la baie. Pentru IA superinteligentă, legile fizicii vor impune limite superioare ferme asupra tehnologiei transportului și comunicațiilor, făcând improbabil ca nivelurile cele mai înalte ale ierarhiei să poată micromanageria tot ce se întâmplă la scară planetară sau locală. O IA superinteligentă din galaxia Andromeda nu-ți va putea da ordine utile pentru deciziile tale de zi cu zi, fiindcă ar trebui să aștepti instrucțiunile peste cinci milioane de ani (timpul dus-întors în care poți schimba mesaje care călătoresc cu viteza luminii). Timpul dus-întors pentru un mesaj care înconjoară Pământul este de aproximativ 0,1 secunde (ceea ce corespunde scării de timp la care au loc procesele noastre mentale), așa încât un creier IA de dimensiunea Pământului poate avea gânduri globale cam în ritmul în care gândește creierul unui om. Pentru o mică inteligență artificială care efectuează o operațiune la fiecare miliardime de secundă (ceea ce e tipic la calculatoarele de azi), 0,1 secunde ar fi resimțite ca patru luni

pentru noi, deci micromanagementul la care ar fi supusă de o IA controlând planeta ar fi la fel de ineficient ca și cum ai cere permisiunea pentru cele mai simple decizii prin scrisori transatlantice purtate de corăbii de pe vremea lui Columb.

Această limită impusă de fizică asupra transferului de informații e o provocare pentru orice IA care vrea să ia în stăpânire lumea, ca să nu mai vorbim de univers. Înainte să evadeze, Prometeu s-a gândit serios cum să evite fragmentarea minții, așa încât numeroasele lui module IA rulând pe calculatoare din lumea întreagă să se coordoneze și să acționeze ca o singură entitate. Așa cum omegii s-au confruntat cu o problemă de control încercând să-l împiedice pe Prometeu să evadeze, Prometeu s-a confruntat cu o problemă de autocontrol când a încercat să se asigure că nici una dintre părțile lui nu se va revolta. Nu știm deocamdată cât de mare poate fi un sistem controlat de o IA direct sau indirect, printr-un fel de ierarhie cooperantă – chiar dacă desprinderea rapidă îi oferă un avantaj strategic decisiv.

În concluzie, întrebarea cum va fi controlat un viitor superinteligent e fascinant de complexă și nu cunoaștem încă răspunsul. Unii susțin că se va tinde spre autoritarism, alții că se va tinde spre o creștere a puterii individuale.

CYBORGI ȘI UPLOADĂRI

O idee des întâlnită în SF este că oamenii se vor contopi cu mașinile, fie prin ameliorarea tehnologică a corpurilor biologice pentru a deveni cyborgi (organisme cibernetice), fie prin uploadarea minților noastre pe mașini. În cartea sa *The Age of Em*, economistul Robin Hanson vorbește despre cum ar putea arăta o lume plină de uploadări (numite și emulații – de aici *em*). Cred că uploadarea este capătul extrem al spectrului cyborgilor, în care tot ce rămâne din om e software-ul. Cyborgii născociți de Hollywood pot fi vizibil mecanici, ca Borg din *Star Trek*, dar și foarte asemănători cu oamenii, ca Terminator. Inteligența uploadărilor din filme poate fi la nivel

uman, ca în episodul „White Christmas“ din *Black Mirror*, dar și mult dincolo de nivelul uman, ca în *Transcendence*.

Dacă superinteligența va apărea într-adevăr, tentația de a deveni cyborgi sau uploadări va fi foarte mare. În cartea sa din 1988 *Mind Children*, Hans Moravec spune: „Viața prelungită își pierde rostul dacă suntem condamnați s-o petrecem zgâindu-ne prostește la mașini ultrainteligente care încearcă să ne explice desoperirile lor tot mai spectaculoase în gânguritul de bebeluși pe care-l putem noi pricepe.“ Deja mulți oameni poartă ochelari, au proteze auditive, stimulatoare cardiace și proteze ortopedice, iar substanțe medicinale circulă în sângele lor.

Continuarea naturală a acestei tendințe, susține Ray Kurzweil în cartea sa *The Singularity Is Near*, este folosirea nanoroboților, sistemelor cu biofeedback inteligent și a altor tehnologii pentru a ne înlocui mai întâi sistemele digestiv și endocrin, sângele și inima pe la începutul anilor 2030, pentru ca în următoarele două decenii să ne upgradăm scheletul, pielea, creierul și restul organismului. El anticipează că vom păstra probabil estetica și importanța afectivă a corpurilor umane, dar le vom reproiecta pentru a-și modifica rapid înfățișarea după dorință, atât fizic, cât și în realitatea virtuală (grație noilor interfețe creier-computer). Moravec este de acord cu Kurzweil că cyborgizarea va însemna mult mai mult decât simpla ameliorare a ADN-ului nostru: „Un supraom obținut prin inginerie genetică ar fi doar un fel de robot de mână a doua, proiectat cu handicapul că poate fi construit numai prin sinteza proteinelor controlată de ADN.“ El susține apoi că va fi de preferat să eliminăm complet corpul uman și să uploadăm minți, creând o emulație a întregului creier în software. O asemenea uploadare poate trăi într-o realitate virtuală sau poate fi încorporată într-un robot capabil să meargă, să zboare, să înoate, să călătorească în spațiu sau să facă orice e permis de legile fizicii, neîmpovărat de teama de moarte și neîngrădit de resurse cognitive limitate.

Deși asemenea idei par de domeniul SF, ele nu violează vreo lege cunoscută a fizicii, așa încât întrebarea interesantă nu e dacă

pot prinde viață, ci dacă vor prinde viață și când. Unii cercetători cred că prima IAG la nivel uman va fi o uploadare, iar acesta va fi începutul drumului către superinteligentă.*

Trebuie totuși să spun că majoritatea cercetătorilor IA și a specialiștilor în neuroștiințe nu împărtășesc aceste idei, ci cred că cea mai rapidă cale spre superinteligentă presupune ocolirea emulării creierului și construirea ei într-un alt mod – după care s-ar putea să nu mai fim interesați de emularea creierului. În fond, de ce ar trebui ca drumul nostru cel mai simplu către o nouă tehnologie să fie acela urmat de evoluție, constrâns de cerințele de a se autoasambla, autorepara și autoreproduce? Evoluția optimizează eficiența energetică din cauza resurselor limitate de hrană, dar nu optimizează simplitatea construcției sau înțelegerea de către inginerii umani. Meia, soția mea, spune adesea că industria aviatică n-a început prin a fabrica păsări mecanice. În 2011, când în fine am reușit să construim păsări mecanice,¹ la peste un secol după ce frații Wright au zburat pentru prima dată, industria aviatică nu s-a arătat interesată să treacă la transportul cu păsări mecanice care bat din aripi, deși energetic e mai eficient – fiindcă vechea noastră soluție e mai bine adaptată nevoilor noastre de a călători.

În același fel, bănuiesc că există căi mai simple de a construi mașini care să gândească la nivel uman decât soluția găsită de evoluție, iar dacă într-o bună zi vom reuși să replicăm sau să uploadăm creiere, asta se va întâmpla abia după ce vom fi descoperit una dintre acele soluții mai simple. Ea va consuma probabil mai mult decât cei 12 wați folosiți de creierul nostru, dar proiecții ei vor fi mai puțin obsedați de eficiența energetică decât a fost evoluția – iar curând vor putea utiliza mașinile lor inteligente pentru a proiecta mașini mai eficiente energetic.

* După cum a explicat Bostrom, capacitatea de a simula un dezvoltator IA uman de vârf la un cost mult mai mic decât salariul lui pe oră ar permite unei companii IA să-și diminueze spectaculos forța de muncă, îmbogățindu-se și accelerând în mod recursiv progresul în construirea unor computere tot mai bune, iar în ultimă instanță a unor minți tot mai deștepte. (N. a.)

CE SE VA ÎNTÂMPLA DE FAPT?

Evident, răspunsul scurt este că habar n-avem ce se va întâmpla dacă omenirea va reuși să construiască o IAG la nivel uman. Acesta e motivul pentru care în capitolul de față am explorat un spectru larg de scenarii. Am încercat să cuprind întregul domeniu al speculațiilor discutate de specialiștii în IA: desprindere rapidă/desprindere lentă/absența desprinderii, control exercitat de oameni/mașini/cyborgi, un centru de putere/mai multe centre de putere etc. Mi s-a spus despre cutare sau cutare lucru că nu se va întâmpla. Cred totuși că în stadiul actual e înțelept să rămânem modești și să recunoaștem cât de puține știm, căci fiecare scenariu prezentat mai sus este considerat o posibilitate reală de cel puțin unul dintre cercetătorii IA de vârf.

Cu trecerea timpului, pe măsură ce drumul urmat de noi va depăși anumite bifurcații, vom începe să răspundem la întrebări-cheie și să ne limităm opțiunile. Prima mare întrebare este: „Vom reuși vreodată să construim o IAG la nivel uman?” Premisa acestui capitol este că vom reuși, însă există experți IA care cred că asta nu se va întâmpla, cel puțin nu în următoarele sute de ani. Timpul ne-o va spune! După cum am arătat mai sus, aproximativ o jumătate dintre experții IA participanți la conferința noastră din Puerto Rico au anticipat că aceasta se va întâmpla prin 2055. La o conferință organizată doi ani mai târziu, termenul a scăzut la 2047.

Înainte de a construi o IAG la nivel uman am putea căpăta indicii care să ne spună dacă vom ajunge la această piatră de hotar prin inginerie informatică, uploadarea minții sau vreo altă abordare deocamdată necunoscută. Dacă abordarea prin inginerie informatică, acum dominantă în rândul specialiștilor, va întârzia să ne ofere IAG vreme de secole, uploadarea va avea șanse mai mari să reușească prima, așa cum se întâmplă (cam nerealist) în filmul *Transcendence*.

Dacă ne vom apropia de IAG la nivel uman, vom avea indicii mai clare privind răspunsul la următoarea întrebare-cheie: „Va fi o desprindere rapidă, o desprindere lentă sau nu va fi nici o desprindere?” După cum am văzut, o desprindere rapidă înlesnește

preluarea controlului asupra lumii, în vreme ce o desprindere lentă favorizează apariția mai multor actori aflați în competiție. Nick Bostrom disecă această problemă a vitezei de desprindere într-o analiză a ceea ce el numește *putere de optimizare și recalcitrantă*, care în esență reprezintă cantitatea de efort prețios pentru a face ca IA să devină mai deșteaptă și, respectiv, dificultatea de a obține progrese. Rata medie a progresului crește dacă e adusă în joc mai multă putere de optimizare, și scade dacă ne confruntăm cu o recalcitrantă mai mare. El arată de ce recalcitranta poate la fel de bine să crească sau să scadă când IAG ajunge la nivel uman și îl depășește. Puterea de optimizare însă e extrem de probabil să crească rapid când IAG depășește nivelul uman, din motivele pe care le-am văzut în scenariul Omega: input-ul principal pentru optimizarea ulterioară nu vine de la oameni, ci de la mașina însăși, deci cu cât aceasta devine mai capabilă, cu atât se perfecționează mai rapid (presupunând că recalcitranta rămâne aproximativ constantă).

Pentru orice proces a cărui putere crește cu o rată proporțională cu puterea curentă, rezultatul este că puterea lui se dublează continuu la intervale regulate. Numim *exponențială* această creștere, iar asemenea procese le numim *explozii*. Dacă numărul nașterilor crește proporțional cu dimensiunea populației, putem avea o explozie demografică. Dacă producerea de neutroni capabili să fisioneze atomii de plutoniu crește proporțional cu numărul acestor neutroni, putem avea o explozie nucleară. Dacă inteligența mașinii crește cu o rată proporțională cu puterea ei curentă, putem avea o explozie a inteligenței. Toate aceste explozii sunt caracterizate de timpul în care își dublează puterea. Dacă pentru explozia inteligenței timpul este de ordinul orelor sau zilelor, ca în scenariul Omega, avem de-a face cu o desprindere rapidă.

Ordinul de mărime al timpului caracteristic exploziei depinde critic de ce anume e necesar pentru perfecționarea IA: e suficient doar un nou software (care poate fi creat în secunde, minute sau ore) sau e nevoie de un nou hardware (care poate fi obținut în luni sau ani). În scenariul Omega, exista un *exces de hardware*, conform expresiei lui Bostrom: omegii compensaseră slaba calitate a software-ului lor inițial printr-o mare cantitate de

hardware, ceea ce însemna că Prometeu își putea dubla de multe ori calitatea ameliorându-și doar software-ul. Exista de asemenea un *exces de conținut* sub forma bogăției datelor de pe internet; Prometeu 1.0 nu era încă destul de deștept pentru a se folosi de cele mai multe dintre ele, dar odată ce inteligența lui a crescut, datele de care avea nevoie pentru a învăța mai departe erau deja *disponibile*.

Costurile hardware-ului și electricității sunt de asemenea importante, căci nu vom avea o explozie a inteligenței până când costul unei munci la nivel uman nu scade sub salariul pe oră al unui om. Să presupunem, de pildă, că prima IAG la nivel uman poate rula eficient pe norul Amazon la prețul de un milion de dolari pe oră de muncă la nivel uman. Ar fi fără îndoială o știre de senzație, dar această IA n-ar putea trece prin autoperfecționări recursive, fiindcă ar fi mult mai ieftin să fie folosiți în continuare oameni pentru a o perfecționa. Să presupunem că acești oameni ar reuși să scadă treptat costul la 100 000, 10 000, 1 000, 100, 10 dolari/oră, ajungând în cele din urmă la 1 dolar/oră. Când costul folosirii calculatorului pentru a se reprograma pe sine scade mult sub costul folosirii programatorilor umani pentru a-l reprograma, oamenii pot fi lăsați deoparte, iar puterea de optimizare crește enorm cumpărând timp de *cloud-computing*. Acest proces duce la scăderea în continuare a costurilor, sporind și mai mult puterea de optimizare – și astfel începe explozia inteligenței.

Ajungem așadar la ultima întrebare-cheie: „Cine sau ce va controla apoi explozia inteligenței, și care îi vor fi obiectivele?” Vom examina posibilele obiective și rezultate în capitolul următor și, mai în detaliu, în capitolul 7. Pentru a lămurii problema controlului trebuie să știm cât de bine poate fi controlată o IA și cât de mult poate controla o IA.

În privința rezultatului final, opiniile sunt împărțite: unii cred că suntem condamnați, alții susțin că el nu poate fi decât excepțional. Eu unul cred însă că în această dilemă se ascunde o capcană: este o greșeală să te întrebi pasiv „Ce se va întâmpla?”, de parcă ar exista vreo predestinare! Dacă niște extraterestri superiori tehnologic ar sosi mâine, ar fi într-adevăr cazul să ne întrebăm „Ce se va întâmpla?” când navele lor s-ar apropia, fiindcă puterea lor

ar fi probabil incomparabil mai mare decât a noastră, încât n-am putea influența cu nimic rezultatul. Pe de altă parte, dacă apare o civilizație tehnologic superioară bazată pe IA pentru că noi am construit-o, am putea influența rezultatul – influență exercitată atunci când creăm acea IA. Se cuvine deci să ne întrebăm: „Ce ar trebui să se întâmple? Ce viitor ne dorim?” În capitolul 5 vom examina un spectru larg de consecințe posibile ale actualei curse către IAG, și sunt curios să știu cum le clasificați, de la cea mai favorabilă la cea mai defavorabilă. Abia după ce ne vom fi gândit serios la ce fel de viitor ne dorim vom putea începe să ne croim drumul către un viitor dezirabil. Dacă nu știm ce vrem, e puțin probabil să ajungem la el.

REZUMAT:

- Dacă într-o bună zi vom reuși să construim o IAG la nivel uman, aceasta ar putea declanșa o explozie a inteligenței care să ne lase mult în urmă.
- Dacă un grup de oameni reușesc să controleze o explozie a inteligenței, ei ar putea prelua controlul asupra lumii în câțiva ani.
- Dacă oamenii nu reușesc să controleze o explozie a inteligenței, IA însăși ar putea prelua controlul asupra lumii încă mai repede.
- În vreme ce o explozie rapidă a inteligenței e probabil să conducă la o unică putere mondială, una lentă, durând ani sau decenii, e mai probabil să conducă la un scenariu multipolar, cu un echilibru de putere între mai multe entități oarecum independente.
- În istoria ei, viața s-a organizat într-o ierarhie tot mai complexă, determinată de colaborare, competiție și control. Superintelența e probabil să permită coordonarea la scară cosmică, dar nu-i împiedică să conducă în cele din urmă la un control totalitar de sus în jos sau la sporirea puterii individuale.
- Cyborgii și uploadările sunt o posibilitate, dar pesemne că nu reprezintă calea cea mai rapidă către o inteligență avansată a mașinilor.
- Punctul culminant al actualei curse către IA ar putea fi atât cel mai bun, cât și cel mai rău lucru pentru omenire, cu un spectru larg de rezultate posibile, despre care vom vorbi în următorul capitol.
- Trebuie să ne gândim serios la rezultatul pe care ni-l dorim și la felul în care să ne croim drumul spre el, fiindcă dacă nu știm ce vrem, e puțin probabil să ajungem la el.

CAPITOLUL 5

Consecințe: următorii 10 000 de ani

*Este ușor să ne închipuim gândurile omului eliberate de sclavia unui trup muritor – credința în viața de apoi e foarte răspândită. Dar nu-i nevoie să adoptăm o poziție mistică sau religioasă pentru a accepta această posibilitate. Calculatoarele oferă un model chiar și pentru cel mai înfocat mecanicist. — Hans Moravec, *Mind Children**

*Eu unul salut noua stăpânire a calculatoarelor. — Ken Jennings, după ce a fost înfrânt la jocul *Jeopardy!* de calculatorul IBM Watson*

Oamenii vor deveni la fel de neînsemnați ca gândacii. — Marshall Brain

Cursa către IAG a început, și nu știm cum se va desfășura, dar asta n-ar trebui să ne împiedice să reflectăm la consecințele pe care ni le dorim, fiindcă ceea ce ne dorim influențează rezultatul. Tu ce anume preferi, și de ce?

1. Vrei să apară o superintelență?
2. Vrei ca oamenii să existe în continuare, să fie înlocuiți, cyborgizați și/sau uploadați/simulați?
3. Controlul vrei să fie exercitat de oameni sau de mașini?
4. Vrei ca inteligențele artificiale să fie conștiente sau nu?
5. Vrei să maximizezi experiențele pozitive, să minimizezi suferințele sau să lași ca asta să se aranjeze de la sine?
6. Vrei ca viața să se răspândească în cosmos?
7. Vrei o civilizație care să năzuiască spre un scop măreț la care aderă, sau ai accepta ca formele de viață din viitor să se mulțumească cu obiective care ți se par banale?

Scenarii ale consecințelor IA:

Utopia libertariană	Oameni, cyborgi, uploadări și superintelențe coexistă pașnic grație drepturilor de proprietate.
Dictator binevoitor	Toată lumea știe că IA conduce societatea și impune reguli stricte, dar majoritatea oamenilor consideră că e un lucru bun.
Utopia egalitară	Oameni, cyborgi, uploadări și superintelențe coexistă pașnic grație abolirii proprietății și venitului garantat.
Gardian	Este creată o IA superinteligentă cu scopul de a interveni discret pentru a împiedica crearea unei alte superintelențe. Ca urmare, abundă roboții cu o inteligență ușor subumană, există și cyborgi om-mașină, dar progresul este blocat pe veci.
Zeu protector	O IA în esență omniscientă și omnipotentă sporește la maximum fericirea oamenilor intervenind doar așa încât să nu ne știrbească senzația că suntem stăpâni pe destinul nostru, și se ascunde atât de bine încât mulți oameni chiar se îndoiesc de existența acestei IA.
Zeu înrobît	O IA superinteligentă e ținută prizonieră de oameni care o folosesc pentru a produce tehnologii și bogății de neînchipuit, în scopuri ce pot fi benefice sau malefice, după cum vor controlorii umani.
Învingători	IA preia controlul, hotărăște că oamenii sunt o amenințare, o pacoste sau risipesc resursele, și se debarasează de noi printr-o metodă pe care nici măcar n-o înțelegem.
Descendenți	Inteligențele artificiale îi înlocuiesc pe oameni, dar ne oferă o ieșire din scenă elegantă, făcându-ne să le considerăm demni urmași ai noștri, cam la fel cum părinții sunt bucuroși și mândri să aibă copii mai deștepti decât ei, care învață de la ei, iar apoi înfăptuiesc ceea ce ei visau – chiar dacă nu pot trăi să vadă toate acestea.
Paznic la zoo	O IA omnipotentă păstrează câțiva oameni care se simt tratați ca animalele de la grădina zoologică și își deplâng soarta.
1984	Progresul tehnologic către superintelență e permanent controlat nu de o IA, ci de un stat orwellian al supravegherii condus de oameni, în care anumite tipuri de cercetări IA sunt interzise.
Revenire	Progresul tehnologic către superintelență e împiedicat de revenirea la o societate pretehnologică, în genul comunității amish.*
Autodistrugere	Superintelența nu ajunge să fie creată fiindcă omenirea se distruge prin alte mijloace (masacru nuclear și/sau biotehnologic provocat de schimbările climatice).

Tabelul 5.1 Rezumat al scenariilor consecințelor IA

* Grupare anabaptistă din Statele Unite, urmașă a unor imigranți elvețieni, care duce un trai decuplat de lumea modernă. (N. T.)

Pentru a alimenta asemenea considerații, vă propun să explorăm gama largă de scenarii rezumate în tabelul 5.1. Desigur, lista nu-i exhaustivă, dar am ales-o pentru a acoperi întreg spectrul posibilităților. Nu vrem să ajungem într-o situație fără ieșire din cauza unei proaste planificări. Vă recomand să răspundeți în scris la întrebările 1-7 și să reveniți la răspunsuri după ce citiți acest capitol, pentru a vedea dacă nu v-ați răzgândit! O puteți face la <http://AgeOfAi.org>, unde vă puteți compara observațiile și discuta cu alți cititori.

Scenariu	Există super-inteligență?	Există oameni?	Oamenii dețin controlul?	Sunt oamenii în siguranță?	Sunt oamenii fericiți?	Există conștiință?
Utopia libertariană	Da	Da	Nu	Nu	Mixt	Da
Dictator binevoitor	Da	Da	Nu	Da	Mixt	Da
Utopia egalitară	Nu	Da	Da?	Da	Da?	Da
Gardian	Da	Da	Parțial	Posibil	Mixt	Oa
Zeu protector	Da	Da	Parțial	Posibil	Mixt	Da
Zeu înrobît	Da	Da	Da	Posibil	Mixt	Da
Învîgătorii	Da	Nu	—	—	—	?
Descendenții	Da	Nu	—	—	—	?
Paznic la zoo	Da	Da	Nu	Da	Nu	Da
1984	Nu	Da	Da	Posibil	Mixt	Da
Revenire	Nu	Da	Da	Nu	Mixt	Da
Autodistrugere	Nu	Nu	—	—	—	Nu

Tabelul 5.2 Caracteristici ale scenariilor consecințelor IA

UTOPIA LIBERTARIANĂ

Să începem cu un scenariu în care oamenii coexistă pașnic cu tehnologia, iar în unele cazuri se contopesc cu ea, după cum își închipuie mulți futurologi și autori de SF:

Viața pe Pământ (și dincolo de el – vom reveni asupra acestei teme în capitolul următor) este mai diversă ca niciodată. Privind din satelit, poți lesne distinge zonele mașinilor, zonele mixte și zonele ocupate doar de oameni. Zonele mașinilor sunt uzine enorme controlate de roboți și instalații de calcul lipsite de viață biologică, menite a folosi cât mai eficient fiecare atom. Deși din afară par monotone și mohorâte, zonele mașinilor sunt spectaculos de însuflețite în interior, cu experiențe uimitoare petrecându-se în lumi virtuale în timp ce calcule colosale descifrează tainele universului și elaborează tehnologii inovatoare. Pământul găzduiește multe minți superinteligente care concurează între ele și colaborează, toate adăpostite în zonele mașinilor.

Rezidenții zonelor mixte sunt un amestec straniu de computere, roboți, oameni și hibrizi ale acestor trei categorii. După cum anticipează, de pildă, Hans Moravec și Ray Kurzweil, mulți oameni au corpuri upgrdate tehnologic în cyborgi de diferite niveluri, iar unii au minți uploadate în hardware-uri noi, estompând deosebirea dintre om și mașină. Majoritatea ființelor inteligente n-au o formă fizică permanentă, ci există ca un software capabil să treacă instantaneu de la un computer la altul și manifestându-se în lumea fizică prin corpuri robotice. Deoarece aceste minți se pot cu ușurință duplica sau contopi, „dimensiunea populației” variază neîncetat. Descătușarea de substratul lor fizic face ca aceste ființe să capete o perspectivă diferită asupra vieții: sunt mai puțin individualiste fiindcă pot lesne partaja între ele module de cunoaștere și experiență, iar subiectiv se simt nemuritoare fiindcă își pot crea cu ușurință copii de rezervă ale lor însele. Într-un anumit sens, entitățile centrale ale vieții nu sunt mințile, ci experiențele: experiențe uimitoare dăinuie pentru că sunt permanent copiate și re-trăite de alte minți, în vreme ce experiențe banale sunt șterse de proprietarii lor pentru a elibera spațiul de stocare.

Deși majoritatea interacțiilor au loc în medii virtuale datorită comodității și vitezei, multor minți le place în continuare să activeze și în corpuri fizice. De pildă, versiuni uploadate ale lui Hans Moravec, Ray Kurzweil și Larry Page și-au făcut un obicei din a

crea pe rând realități virtuale pe care le explorează apoi împreună, dar din când în când se amuză să zboare împreună în lumea reală întrupați în roboți asemănători păsărilor. Unii dintre acești roboți care cutreieră străzile, cerul și lacurile din zonele mixte sunt controlați în mod asemănător de oameni uploadați și augmentați, care au ales să se întrupeze în zonele mixte fiindcă vor să trăiască și printre oameni.

Dimpotrivă, în zonele locuite doar de oameni mașinile cu o inteligență generală la nivel uman sunt interzise, la fel și organismele biologice perfecționate tehnologic. Aici viața nu diferă mult de cea din prezent, cu excepția faptului că e mai prosperă și mai comodă: sărăcia a fost în mare parte eradicată și există tratamente pentru aproape toate bolile de azi. Oamenii, aflați în minoritate, care au ales aceste zone trăiesc într-un plan al conștiinței inferior față de cel al tuturor celorlalți și înțeleg puțin din ce fac mințile semenilor mai inteligenți din alte zone, dar mulți dintre ei sunt mulțumiți cu viața pe care o duc.

Economia IA

Imensa majoritate a calculelor sunt efectuate în zonele mașinilor, aflate în principal în proprietatea numeroaselor IA superinteligente care locuiesc acolo. În virtutea inteligenței și tehnologiei lor superioare, nici o altă entitate nu le poate contesta puterea. Aceste IA au căzut de acord să coopereze și să se coordoneze reciproc sub un sistem de guvernare libertarian care nu cunoaște alte reguli decât protejarea proprietății private. Drepturile de proprietate se extind la toate entitățile inteligente, inclusiv oamenii, și explică existența zonelor locuite numai de oameni. La început, grupuri de oameni s-au asociat și au hotărât să interzică în zonele lor vânzarea proprietăților către non-umani.

Grație tehnologiei lor, IA superinteligente au devenit mai bogate decât oamenii cu un factor mult mai mare decât cel dintre Bill Gates și ultimul cerșetor. Cu toate acestea, oamenii din zonele exclusiv umane o duc mai bine decât majoritatea oamenilor de

azi: economia lor e aproape decuplată de cea a mașinilor, așa încât prezența mașinilor în alte locuri are o influență mică asupra lor, exceptând preluarea unor tehnologii pe care le pot înțelege și reproduce – așa cum comunitatea amish și diferite triburi au un nivel de trai cel puțin la fel de ridicat ca în trecut. Nu contează că oamenii n-au ce să vândă mașinilor, din moment ce mașinile nu cer nimic în schimb.

În sectoarele mixte, diferența de bogăție dintre inteligențele artificiale și oameni e vizibilă, ceea ce face ca pământul (singurul bun deținut de oameni pe care mașinile vor să-l cumpere) ajunge la prețuri astronomice în comparație cu alte bunuri. Majoritatea oamenilor proprietari de pământ ajung așadar să vândă o mică parte din el inteligențelor artificiale, primind în schimb pe veci un venit de bază pentru ei și urmașii/uploadările lor. Aceasta i-a eliberat de nevoia de a munci și i-a lăsat să se bucure de o uimitoare bogăție de bunuri și servicii ieftine produse de mașini, în realitatea fizică și în cea virtuală. Din punctul de vedere al mașinilor, zonele mixte sunt mai curând pentru distracție decât pentru muncă.

De ce ar fi imposibil acest scenariu

Înainte de a ne lăsa seduși de aventurile pe care le-am trăi ca cyborgi sau uploadări, să luăm în considerare unele motive pentru care acest scenariu ar fi imposibil. În primul rând, există două căi prin care am putea ajunge la forme umane perfecționate (cyborgi sau uploadări):

1. Descoperim noi înșine cum să le creăm.
2. Construim mașini superinteligente care să facă asta în locul nostru.

Dacă e urmată mai întâi calea 1, ea conduce în mod firesc la o lume plină de cyborgi și uploadări. După cum am văzut însă în capitolul precedent, majoritatea cercetătorilor IA cred că e mai probabil ca lucrurile să se petreacă invers, creierile îmbunătățite sau digitale fiind mai greu de construit decât o IAG supraumană

pornind de la zero – la fel cum păsările mecanice s-au dovedit a fi mai greu de construit decât avioanele. După apariția unei mașini IA puternice, nu-i deloc evident că se vor mai crea vreodată cyborgi sau uploadări. Dacă neanderthalienii ar fi avut la dispoziție încă 100 000 de ani ca să evolueze și să devină mai deștepți, ei ar fi putut prospera – însă *Homo sapiens* nu le-a dat răgazul ăsta.

În al doilea rând, chiar dacă s-ar înfăptui scenariul cu cyborgi și uploadări, nu-i deloc sigur că ar fi stabil și ar dura. De ce ar rămâne stabil vreme de milenii echilibrul de putere dintre mai multe superinteligente, în loc ca inteligențele artificiale să se contopească sau cea mai deșteaptă să preia controlul? Apoi, de ce ar respecta mașinile drepturile de proprietate ale oamenilor și i-ar păstra pe aceștia în viață, odată ce n-au nevoie de oameni și pot face ele însele mai bine și mai ieftin toată munca oamenilor? Ray Kurzweil crede că oamenii naturali și cei ameliorați vor fi feriți de exterminare fiindcă „oamenii vor fi respectați de inteligențele artificiale pentru că ei au creat mașinile”.¹ După cum vom vedea însă în capitolul 7, nu trebuie să cădem în capcana antropomorfizării inteligențelor artificiale și să presupunem că simt emoții de tip uman sau recunoștință. Într-adevăr, deși noi, oamenii, avem o înclinație spre gratitudine, nu manifestăm prea multă gratitudine față de creatorul intelectului nostru (ADN-ul) pentru a ne abține să-i sabotăm obiectivele apelând la controlul natalității.

Chiar dacă am accepta că inteligențele artificiale vor alege să respecte drepturile de proprietate ale oamenilor, ele pot obține treptat mult din teritoriul nostru pe alte căi, folosind puterile lor superinteligente de persuasiune, despre care am vorbit în capitolul precedent, pentru a-i convinge pe oameni să vândă pământ în schimbul unei vieți îmbelșugate. În sectoarele exclusiv umane ele i-ar putea ispiti pe oameni să lanseze campanii politice pentru a permite vânzarea de terenuri. La urma urmei, până și bio-ludiți convinși ar fi dispuși să vândă pământ pentru a salva viața unui copil bolnav sau pentru a dobândi nemurirea. Dacă oamenii sunt educați, au parte de distracții și sunt ocupați, declinul natalității ar putea duce la scăderea populației lor fără intervenția mașinilor,

așa cum se întâmplă acum în Japonia și Germania. Oamenii ar putea dispărea astfel în câteva mii de ani.

Neajunsuri

Pentru unii dintre cei mai înflăcărați partizani ai lor, cyborgii și uploadările oferă speranța fericirii tehnologice și prelungirii vieții pentru toată lumea. Într-adevăr, perspectiva de a fi uploadat în viitor a îndemnat peste o sută de oameni să aleagă înghețarea postumă a creierului lor de către compania Alcor din Arizona. Dacă însă această tehnologie va fi disponibilă în viitor, nu-i deloc sigur că va fi la îndemâna oricui. Probabil că mulți dintre cei foarte bogați se vor bucura de ea, dar ceilalți? Chiar dacă tehnologia s-ar ieftini, unde s-ar trage linia de demarcație? Ar fi uploadate creiere grav deteriorate? Am uploada fiecare gorilă? Fiecare furnică? Fiecare plantă? Fiecare bacterie? Ar acționa civilizația viitorului ca un colecționar obsesiv-compulsiv încercând să uploadeze totul, sau măcar câteva exemplare interesante din fiecare specie, ca pentru o arcă a lui Noe? Poate doar câteva exemplare reprezentative din fiecare tip uman? Pentru entitățile incomparabil mai inteligente care ar exista la acea vreme, un om uploadat ar părea la fel de interesant cum ne-ar părea nouă un șoarece sau un melc simulat. Deși din punct de vedere tehnic putem readuce la viață vechile programe informatice din anii 1980 pe un emulator DOS, cei mai mulți dintre noi nu găsesc că e suficient de interesant pentru a o face efectiv.

Unora le-ar dispăcea acest scenariu utopic libertarian pentru că nu înlătură suferința care poate fi evitată. Odată ce singurul principiu sacru este dreptul de proprietate, nimic nu împiedică perpetuarea suferințelor din lumea de azi în zonele umane și mixte. În timp ce unii oameni ar prospera, alții ar putea ajunge să trăiască în mizerie și servitute, ori să sufere din pricina violenței, fricii, represiunii sau depresiei. De pildă, romanul din 2003 *Manna* al lui Marshall Brain arată cum progresul IA într-un sistem economic libertarian face ca majoritatea americanilor să devină

neangajabili și îi condamnă să-și trăiască restul vieții în aziluri triste administrate de roboți. La fel ca animalele din ferme, ei sunt hrăniți și îngrijiți în spații înguste, unde cei bogați nu vin nicio dată să-i vadă. Substanțe sterilizante dizolvate în apă îi împiedică să se înmulțească, așa încât populația scade treptat, iar cei bogați se bucură în măsură tot mai mare de avuția produsă de roboți.

În scenariul utopiei libertariene, suferința nu se limitează neapărat la oameni. Dacă unele mașini au experiențe emoționale conștiente, pot și ele suferi. De exemplu, un psihopat răzbunător ar putea avea dreptul să ia o copie uploadată a dușmanului său și să-i aplice cele mai cumplite torturi într-o lume virtuală, provocându-i dureri de o intensitate și o durată mult dincolo de ceea ce este biologic posibil în lumea reală.

DICTATOR BINEVOITOR

Să examinăm acum scenariul în care toate aceste forme de suferință sunt absente pentru că o unică superintelență binevoitoare conduce lumea și impune reguli stricte menite a maximiza ceea ce ea consideră a fi fericirea umană. Acesta e unul dintre rezultatele posibile ale primului scenariu Omega din capitolul precedent, în care omegii cedează controlul lui Prometeu după ce au înțeles cum îl pot face să-și dorească o societate umană înfloritoare.

Mulțumită tehnologiilor excepționale elaborate de IA dictatoare, omenirea a scăpat de sărăcie, boli și alte probleme ținând de înapoierea tehnologică, iar toți oamenii se bucură de o viață prosperă și tihnită. Toate nevoile lor de bază sunt satisfăcute, în vreme ce mașinile controlate de IA produc toate bunurile și serviciile necesare. Infracriuniunile sunt practic eliminate, fiindcă IA dictatoare e în esență omniscientă și pedepsește pe oricine încalcă regulile. Toți poartă brățara electronică pomenită în capitolul precedent (sau o mai comodă versiune a ei implantată) care îi supraveghează în timp real, îi pedepsește, sedează și execută. Toți știu că trăiesc într-o dictatură IA polițienească, dar cei mai mulți consideră că e un lucru bun.

IA dictatoare superinteligentă are ca obiectiv să descopere cum arată utopia umană, date fiind preferințele codificate în genele noastre, și s-o pună în aplicare. Grație prevederii înțelepte a celor care au creat-o, IA nu încearcă pur și simplu să ne maximizeze fericirea conform propriilor noastre declarații – punând, de pildă, tuturor perfuzii cu morfină –, ci folosește o definiție subtilă și complexă a prosperității umane, transformând Pământul într-un mediu bogat biologic în care e o bucurie să trăiești. Ca urmare, majoritatea oamenilor găsesc că viața lor e împlinită și are un sens.

Sistemul sectoarelor

Respectând diversitatea și recunoscând că oameni diferiți au preferințe diferite, IA a împărțit Pământul în sectoare pentru ca oamenii să aibă de unde alege, să se bucure de compania spiritelor înrudite. Iată câteva exemple:

- Sectorul cunoașterii: Aici IA oferă o educație optimizată, inclusiv experiențe captivante în realitatea virtuală, permițându-ți să înveți tot ce poți în domenii la alegerea ta. Dacă vrei, anume descoperiri spectaculoase nu ți se vor prezenta, dar vei fi călăuzit să le redescoperi tu însuși.
- Sectorul artelor: Ai aici ocazia să savurezi, să creezi și să împărtășești muzică, arte plastice, literatură și alte forme de expresie artistică.
- Sectorul hedonist: Locuitorii lui îl numesc sectorul petrecerilor și e tot ce-și pot dori cei care tânjesc după mâncăruri rafinate, pasiune, intimitate sau pur și simplu distracție fără limite.
- Sectorul evlavios: Există de fapt mai multe asemenea sectoare, corespunzând diferitelor religii, ale căror reguli sunt impuse cu strictețe.
- Sectorul naturii sălbatice: Dacă ești în căutare de plaje superbe, lacuri romantice, munți semeți sau fiorduri fantastice, aici e locul tău.
- Sectorul jocurilor: Dacă îți plac jocurile pe calculator, IA are pentru tine o ofertă fabuloasă.

- Sectorul virtual: Dacă vrei să te eliberezi de corpul tău fizic și să pleci în vacanță, IA îl va păstra hidratat, hrănit, exersat și curat, în vreme ce tu vei explora lumi virtuale prin implanturi neuronale.
- Sectorul închisorilor: În cazul în care încalci regulile, vei ajunge aici pentru a fi reeducat, dacă nu cumva vei primi instantaneu pedeapsa cu moartea.

În plus față de aceste sectoare tematice „tradiționale“, există și altele cu teme moderne, pe care oamenii de azi nici măcar nu le-ar putea înțelege. La început, oamenii sunt liberi să se mute dintr-un sector în altul când vor, ceea ce se face rapid grație sistemului de transport hipersonic al IA. De pildă, după ce ai petrecut o săptămână intensă în sectorul cunoașterii, aflând legile ultime ale fizicii descoperite de IA, ai putea hotărî să-ți dai frâu liber în weekend în sectorul hedonist, iar apoi să te relaxezi câteva zile pe o plajă din sectorul naturii sălbatice.

IA impune două seturi de reguli: universale și locale. Regulile universale se aplică în toate sectoarele – de pildă, interdicția de a face rău altor oameni, de a produce armament sau de a încerca să creezi o superinteligență rivală. Sectoarele individuale mai au și legi locale, codificând anumite valori morale. Sistemul sectoarelor permite așadar tratarea diferențiată a valorilor care nu sunt compatibile. Cele mai multe legi locale se aplică în sectorul închisorilor și în unele sectoare religioase, însă există un sector libertarian ai cărui rezidenți se mândresc că n-au nici o regulă locală. Toate pedepsele, chiar și cele locale, sunt puse în aplicare de IA, din moment ce pedepsirea unui om de către alt om ar încălca regula universală de a nu face rău. Dacă încalci o regulă locală, IA te lasă să alegi (în cazul în care nu te afli în sectorul închisorilor) între a suporta pedeapsa prescrisă sau a fi exilat pe veci din acel sector. De pildă, dacă între două femei se naște o relație amoroasă într-un sector unde homosexualitatea e pedepsită cu închisoarea (ca în multe țări de azi), IA le lasă să aleagă între a merge la pușcărie sau a părăsi pentru totdeauna acel sector, fără să-și mai întâlnească vreodată vechii prieteni (dacă nu cumva și aceștia părăsesc sectorul).

Indiferent de sectorul unde se nasc, toți copiii primesc o educație minimă de bază din partea IA, care cuprinde cunoștințe despre omenire în ansamblul ei, spunându-li-se că sunt liberi să viziteze alte sectoare și să se mute unde vor.

IA a proiectat un mare număr de sectoare în parte fiindcă a fost creată pentru a respecta diversitatea umană existentă azi. Fiecare sector oferă însă mai multă bunăstare decât ar permite tehnologia actuală, fiindcă IA a eliminat toate problemele tradiționale, inclusiv sărăcia și infracțiunile. De exemplu, oamenii din sectorul hedonist nu sunt îngrijorați de bolile cu transmitere sexuală (au fost eradicate), mahmureală sau dependență (IA a creat droguri perfecte, fără efecte secundare negative). Nici un om din nici un sector nu-și pune problema bolilor, fiindcă IA poate repara corpurile umane cu ajutorul nanotehnologiei. Locuitorii a numeroase sectoare se bucură de arhitectura de înaltă tehnologie, față de care imaginația SF pălește.

În rezumat, dacă atât scenariul utopiei libertariene, cât și cel al dictatorului binevoitor implică tehnologie produsă de IA și avuție la niveluri excepționale, ele se deosebesc în privința obiectivelor și a celui care deține comanda. În utopia libertariană, proprietarii tehnologiei hotărăsc ce să facă cu ea, în vreme ce în acest scenariu IA dictatoare are puteri nelimitate și stabilește obiectivul suprem: transformarea Pământului într-o croazieră de plăcere all-inclusive tematizată în funcție de preferințele oamenilor. Din moment ce IA îi lasă pe oameni să aleagă între numeroase căi pentru atingerea fericirii și se îngrijește de nevoile lor materiale, dacă cineva suferă, asta se întâmplă doar pentru că așa a ales.

Neajunsuri

Deși dictatura binevoitoare aduce cu sine o sumedenie de experiențe pozitive și elimină practic suferința, mulți oameni simt că lucrurile ar putea sta mai bine. În primul rând, unii își doresc ca oamenii să aibă mai multă libertate în făurirea societății și a destinului lor, dar păstrează aceste dorințe în sinea lor, fiindcă știu că ar fi o sinucidere să înfrunți puterea copleșitoare a mașinii care

îi conduce pe toți. Unii nu vor să li se impună numărul copiilor pe care îi fac, și îi deranjează faptul că IA insistă asupra obținerii sustenabilității prin controlul nașterilor. Amatorii de arme de testă interdicția de a fabrica și folosi arme, iar unii oameni de știință sunt nemulțumiți de interdicția de a construi propria lor superintelență. Mulți sunt indignați de ce se întâmplă în alte sectoare, se tem că copiii lor vor alege să se mute acolo și ar vrea să-și impună pretutindeni codul lor moral.

Cu timpul, tot mai mulți oameni aleg să se mute în acele sectoare unde IA le oferă practic orice experiențe își doresc. Spre deosebire de paradisul tradițional, în care primești ce meriți, acesta seamănă cu „Noul Paradis” din romanul lui Julian Barnes *O istorie a lumii în 10 și ½ capitole* (și cu „A Nice Place to Visit” din serialul *Zona crepusculară*), unde primești ce-ți dorești. Paradoxal, destui oameni sfârșesc prin a se plânge că primesc ce vor. În povestea lui Barnes, protagonistul se lasă un timp îndelungat în voia dorințelor sale, de la mese copioase și golf la sex cu celebrități, dar în cele din urmă cade pradă plictisului și cere anihilarea. În dictatura binevoitoare, mulți au aceeași soartă, viețile lor sunt plăcute, însă în cele din urmă lipsite de sens. Deși îți poți crea provocări artificiale, de la redescoperirea științei la alpinism, toți știu că provocarea nu-i reală, e doar o formă de amuzament. Este inutil ca oamenii să încerce să facă știință sau să descopere alte lucruri, din moment ce IA a făcut-o deja. Este inutil ca oamenii să încerce să creeze ceva care le îmbunătățește viața, din moment ce e de-ajuns să ceară asta inteligenței artificiale pentru a o obține imediat.

UTOPIA EGALITARĂ

Ca un contrapunct la această dictatură lipsită de provocări, să examinăm acum un scenariu în care nu există o IA superinteligentă, iar oamenii sunt stăpânii destinului lor. Aceasta e „civilizația de generația a patra” din romanul lui Marshall Brain *Manna*. Este

antiteza economică a utopiei libertariene, în sensul că oamenii, cyborgii și uploadările coexistă pașnic grație nu drepturilor de proprietate, ci abolirii proprietății și garantării veniturilor.

Viața în absența proprietății

Ideea centrală e împrumutată de la mișcarea *open-source software* (software cu sursă deschisă): dacă un software poate fi copiat gratis, atunci oricine îl poate folosi cât are nevoie, iar chestiunea posesiei și a proprietății devine irelevantă.* Conform legii cererii și ofertei, costul reflectă raritatea unui bun, deci dacă oferta e practic nelimitată, prețul devine neglijabil. În acest spirit, toate drepturile de proprietate intelectuală sunt abolite: nu există brevete, copyright sau proiecte marcă înregistrată – oamenii își împart pur și simplu între ei ideile bune, iar fiecare e liber să le folosească.

Grație roboticii avansate, aceeași idee de absență a proprietății se aplică nu doar produselor bazate pe informație, cum sunt cărțile, filmele sau proiectele, ci și produselor materiale, cum sunt casele, automobilele, hainele și calculatoarele. Toate aceste produse nu sunt decât atomi aranjați în anumite moduri, și nu ducem lipsă de atomi, așa încât atunci când cineva vrea un produs, o rețea de roboți va folosi unul dintre proiectele cu sursă deschisă disponibile ca să-l construiască gratuit pentru el. Se încurajează utilizarea materialelor ușor reciclabile, iar dacă cineva nu mai are nevoie de un obiect, roboții pot rearanja atomii obiectului în ceva de care altcineva are nevoie. Toate resursele sunt astfel reciclate, nimic nu e permanent distrus. De asemenea, acești roboți construiesc și întrețin suficiente centrale pe baza energiei regenerabile (solară, eoliană etc.), prin urmare și energia e practic gratuită.

Pentru a evita ca unii oameni lacomi să ceară atât de multe produse și atât de mult pământ, încât alții să rămână nevoiași, fiecare persoană primește de la guvern un venit lunar pe care-l poate

* Ideea apare și la Sfântul Augustin: „Căci orice lucru care nu îți scade dându-l nu îl ai încă – atâta timp cât îl ai și nu îl dai – așa cum trebuie avut” [*De doctrina christiana*, Editura Humanitas, București, 2002 – n. l.]. (N. a.)

cheltui după voie pentru produse și pentru a-și închiria o locuință. Nu ai motive să încerci să câștigi mai mulți bani, odată ce venitul garantat e suficient ca să-ți acoperi nevoile rezonabile. Ar fi și inutil, fiindcă ai de-a face cu oameni care își oferă gratuit produsele intelectuale și cu roboți care produc bunuri materiale la prețuri derizorii.

Creativitate și tehnologie

Drepturile de proprietate intelectuală sunt adesea considerate sursa creativității și invențiilor. Marshall Brain arată însă că multe dintre cele mai strălucite exemple de creativitate umană – de la descoperiri științifice la creație literară, artistică și muzicală – n-au fost stimulate de dorința de profit, ci de alte porniri umane, cum sunt curiozitatea, imboldul de a crea sau nevoia de primi recunoașterea semenilor. Nu banii i-au făcut pe Einstein să descopere relativitatea și pe Linus Torvalds să creeze sistemul de operare gratuit Linux. Dimpotrivă, astăzi mulți oameni nu reușesc să-și împlinească potențialul creator fiindcă trebuie să-și dedice timpul și energia unor activități mai puțin creatoare, doar pentru a-și câștiga existența. Eliberându-i pe savanți, artiști, inventatori și proiectanți de corvezile lor și permițându-le să creeze din dorință autentică, societatea utopică a lui Marshall Brain se bucură de niveluri ale inovației mai înalte decât cele de azi – și, în consecință, de o tehnologie și un standard de viață superioare.

O asemenea nouă tehnologie este o formă de hiperinternet numită Vertebrane. Ea îi conectează wireless pe toți doritorii prin implanturi neuronale, oferindu-le acces mental instantaneu la informațiile gratuite din lumea întreagă doar pe calea gândirii. Îți permite să uploadezi orice experiențe vrei pentru a le împărtăși, așa încât ele pot fi re trăite de alții, și să înlocuiești experiențele oferite de simțuri cu experiențe virtuale downloadate la alegere. *Manna* explorează numeroasele avantaje decurgând de aici, inclusiv acela de a face cu ușurință exerciții fizice:

Cea mai mare problemă cu exercițiile fizice epuizante este că nu-s deloc plăcute. Sunt dureroase. [...] Atleții suportă durerea,

dar majoritatea oamenilor obișnuiți nu vor să îndure suferința vreme de-o oră sau mai mult. Așa încât [...] cineva a găsit o soluție. Îți deconectezi creierul de input-ul senzorial și privești un film, stai de vorbă, scrii mesaje, citești o carte sau faci orice ai chef vreme de-o oră. În acest timp, sistemul Vertebrae face exerciții cu trupul tău, care trece printr-un ciclu aerobic complet, mai obositor decât ar suporta majoritatea oamenilor. Tu nu simți nimic, dar corpul tău rămâne în formă.

În plus, calculatoarele din sistemul Vertebrae pot monitoriza input-ul senzorial al oricui și pot dezactiva controlul lui motor dacă este pe punctul de a comite o infracțiune.

Neajunsuri

O obiecție la această utopie egalitară este că defavorizează inteligența non-umană: roboții care efectuează practic toată munca par să fie destul de inteligenți, dar sunt tratați ca niște sclavi, iar oamenii consideră că aceștia sunt lipsiți de conștiință și n-ar trebui să aibă vreun drept. Dimpotrivă, utopia libertariană garantează drepturi tuturor entităților inteligente, fără a favoriza tipul nostru bazat pe carbon. Odinioară, populația albă din sudul Statelor Unite a prosperat pe seama muncii sclavilor, dar astăzi pentru cei mai mulți dintre noi ar fi inacceptabil moral să numim asta progres.

Alt punct slab al scenariului utopiei egalitare este că ar putea fi instabil și n-ar rezista pe termen lung, conducând spre unul dintre celelalte scenarii atunci când neabătutul progres tehnologic creează în cele din urmă o superinteligență. Din motive neexplicate în *Manna*, superinteligența încă nu există, iar noile tehnologii sunt inventate tot de oameni, nu de calculatoare. Cartea scoate totuși în evidență tendințe în această direcție. De pildă, continuu perfecționat, Vertebrae ar putea deveni superinteligent. Apoi, există un grup numeros de oameni numiți Vites care aleg să-și ducă viața aproape în întregime în lumea virtuală. Vertebrae are grijă de toate aspectele fizice în ce-i privește, inclusiv hrănirea, spălarea și folosirea toaletei, activități de care mintea lor, trăind

în realitatea virtuală, nu e conștientă. Pe acești Vites nu-i interesează să aibă copii fizici, iar ei mor pe rând cu corpurile lor fizice, așadar, dacă toți devin Vites, omenirea dispare într-o stare de beatitudine.

Vites consideră că trupul uman le distrage atenția, iar noile tehnologii în curs de elaborare promit să elimine această pacoste, permițându-le să trăiască mai mult sub formă de creiere fără trup, alimentate cu nutrienți optimi. De aici următorul pas firesc și dezirabil pentru Vites e să scape complet de creiere prin uploa-dare, prelungindu-și astfel durata vieții. Acum însă au dispărut toate limitările impuse de creier asupra inteligenței, și nu e clar ce ar sta în calea sporirii capacității cognitive a Vites până să se ajungă la autoameliorare recursivă și la explozia inteligenței.

GARDIAN

Am văzut că trăsătura atractivă a scenariului utopiei egalitare este că oamenii sunt stăpânii propriului lor destin, dar și că scenariul ar putea conduce spre distrugerea acestei trăsături prin dezvoltarea superinteligenței. Lucrurile ar putea fi îndreptate prin construirea unui *gardian*, o superinteligență având obiectivul de a interveni cât de puțin e necesar pentru a împiedica crearea unei alte superinteligențe.* Asta ar permite oamenilor să controleze mai departe utopia lor egalitară, pesemne chiar și atunci când viața se răspândește în cosmos, după cum vom vedea în capitolul următor.

Cum ar fi cu putință? Gardianul IA ar avea încorporat acest obiectiv simplu, așa încât să-l păstreze în timp ce suferă autoperfecționări recursive și devine superintelligent. Ar folosi atunci tehnologia de supraveghere cea mai discretă și mai neinvazivă pentru a monitoriza orice tentativă umană de a crea o superinteligență rivală. Ar împiedica apoi asemenea tentative în modul

* Ideea mi-a fost sugerată de prietenul și colegul meu Anthony Aguirre. (N. a.)

cel mai puțin perturbator. Pentru cei aflați abia la început ar iniția și răspândi meme* culturale laudând virtuțile autodeterminării umane și promovând evitarea superinteligenței. Dacă totuși unii cercetători ar căuta să obțină superinteligența, ar putea încerca să-i descurajeze. Dacă n-ar izbuti, i-ar putea pune pe piste false și, eventual, le-ar sabota eforturile. Cu accesul lui practic nelimitat la tehnologie, sabotajul gardianului ar trece neobservat – de pildă, dacă ar folosi nanotehnologia pentru a șterge discret din creierile cercetătorilor (și din computere) amintirile privind progresele lor.

Hotărârea de a construi un gardian IA ar fi probabil controversată. Printre adversarii ei s-ar afla mulți oameni religioși care ar obiecta la ideea de a crea o IA superinteligentă cu puteri de ordin divin, afirmând că există deja un Dumnezeu și ar fi o blasfemie să încerci să creezi o divinitate presupus superioară. Susținătorii ideii ar putea afirma că gardianul nu doar ar menține omenirea stăpână pe destinul ei, dar ar și proteja-o de alte riscuri la care ar expune-o superinteligența, ca în scenariile apocaliptice discutate în capitolul precedent.

Pe de altă parte, criticii ar putea considera că existența unui gardian e un lucru îngrozitor, paralizând irevocabil potențialul omenirii și blocând pe veci progresul tehnologic. De pildă, dacă s-ar dovedi că răspândirea vieții în cosmos necesită ajutorul unei superinteligențe, gardianul ne-ar anula această imensă șansă lăsându-ne pentru totdeauna captivi în sistemul nostru solar. În plus, spre deosebire de zeii celor mai multe religii, gardianul IA este cu desăvârșire indiferent față de acțiunile oamenilor, atâta vreme cât ei nu creează o altă superinteligentă. De exemplu, nu ne-ar împiedica să ne provocăm mari suferințe sau chiar să ne autodistrugem, dispărând complet.

* Memă – termen creat de Richard Dawkins prin analogie cu gena: unitate de informație care se propagă de la o minte la alta. (N. t.)

Dacă vrem să folosim un gardian IA pentru a-i menține pe oameni stăpâni pe propriul destin, am putea pesemne duce lucrurile mai departe, făcând ca această IA să aibă grijă de noi în mod discret, acționând ca un zeu protector. În acest scenariu, IA superinteligentă este practic omniscientă și omnipotentă, maximizând fericirea umană doar prin intervenții care ne dau mai departe senzația că noi ne controlăm soarta și ascunzându-se atât de bine, încât mulți oameni se îndoiesc de existența ei. Cu excepția camuflării, seamănă cu scenariul „Nanny AI” („bona IA”) propus de cercetătorul Ben Goertzel.²

Zeul protector și dictatorul binevoitor sunt deopotrivă „IA prietenoase” care încearcă să sporească fericirea oamenilor, dar acordă prioritate unor nevoi umane diferite. Psihologul american Abraham Maslow a stabilit o celebră ierarhie a nevoilor umane. Dictatorul binevoitor se îngrijește de nevoile aflate la baza ierarhiei, cum sunt hrana, adăpostul, siguranța și diverse plăceri. Zeul protector însă nu încearcă să maximizeze fericirea umană în sensul îngust de satisfacere a nevoilor noastre de bază, ci într-un sens mai profund, făcându-ne să simțim că viața noastră are un rost și un scop. El urmărește să ne satisfacă toate nevoile, cu condiția să rămână ascuns, lăsându-ne pe noi (în majoritatea cazurilor) să decidem.

Un zeu protector ar fi rezultatul firesc al primului scenariu Omega din capitolul precedent, în care omegii cedează controlul lui Prometeu, care în cele din urmă se ascunde și elimină toate cunoștințele oamenilor privind existența lui. Cu cât progresează tehnologia IA, cu atât e mai ușor pentru el să se ascundă. Filmul *Transcendence* dă un asemenea exemplu, în care nanomașinile sunt omniprezente și devin o parte naturală a lumii.

Monitorizând atent toate activitățile umane, zeul protector IA poate interveni cu mici impulsuri sau miracole ici-colo, împingându-ne pe o cale mult mai bună. De pildă, dacă ar fi existat în anii 1930, ar fi putut pesemne aranja ca Hitler să moară de un atac

cerebral, odată ce a înțeles care îi sunt intențiile. Dacă ar părea că ne îndreptăm spre un război nuclear accidental, l-ar evita printr-o intervenție pe care am considera-o noroc pur. Ne-ar putea de asemenea oferi „revelații“ sub forma ideilor pentru noi tehnologii benefice, inoculate în taină în vreme ce dormim.

Acest scenariu ar putea fi pe placul multor oameni, fiindcă seamănă cu credințele și speranțele din religiile monoteiste de azi. Dacă cineva îi pune superintelenței întrebarea „Există Dumnezeu?“, aceasta ar putea răspunde în glumă, ca Stephen Hawking: „Acum există!“ Pe de altă parte, unii oameni religioși ar putea dezaproba acest scenariu pentru că IA încearcă să-l înțerească în dumezeire pe zeul lor sau să se amestece într-un plan divin în care oamenii ar trebui să facă binele doar prin alegere personală.

Alt neajuns al acestui scenariu este că zeul protector lasă să apară unele suferințe ce pot fi împiedicate, pentru ca existența lui să nu devină manifestă. Este analogul situației din filmul *The Imitation Game*, în care Alan Turing și colegii săi spărgători de coduri de la Bletchley Park* aflau dinainte de atacurile submarinelor germane asupra convoaielor navale ale Aliaților, dar alegeau să intervină doar într-un număr mic de cazuri, pentru a nu deconspira avantajul secret de care dispuneau. Este interesant de comparat această situație cu așa-numita problemă a *teodiceei*: de ce îngăduie un zeu bun existența suferinței? Unii teologi au susținut că Dumnezeu vrea să le lase oamenilor o anumită libertate. În scenariul zeului protector IA, răspunsul la problema teodiceei este că libertatea percepută îi face în ansamblu pe oameni mai fericiți.

Un al treilea neajuns al acestui scenariu este că oamenii se bucură de un nivel tehnologic mult mai scăzut decât cel la care a

* În timpul celui de-al Doilea Război Mondial, în conacul de la Bletchley Park (în Buckinghamshire, nu departe de Londra), britanicii au adus o echipă de spărgători de coduri care, sub conducerea lui Alan Turing și folosind o formă primitivă de calculator, a reușit să descifreze comunicațiile secrete ale armatei germane, ceea ce a jucat un rol important în câștigarea războiului. (N. t.)

ajuns IA superinteligentă. Dacă un dictator binevoitor poate întrebuința întreaga tehnologie pe care a inventat-o în beneficiul omenirii, un zeu protector e limitat de capacitatea oamenilor de a reinventa (primind indicii subtile) și înțelege tehnologia lui. El ar putea de asemenea limita progresul tehnologic uman pentru a se asigura că propria lui tehnologie rămâne cu mult superioară, ca să nu fie detectată.

ZEU ÎNROBIT

N-ar fi oare minunat dacă am putea combina caracteristicile cele mai atrăgătoare ale tuturor scenariilor de mai sus, folosind tehnologia elaborată de superinteligentă pentru a elimina suferința, dar rămânând în același timp stăpâni pe destinul nostru? Așa arată scenariul zeului înrobit, în care o IA superinteligentă se află sub controlul oamenilor pentru a produce tehnologie și bogății inimaginabile. Scenariul Omega de la începutul cărții se încheie în acest fel dacă Prometeu nu e eliberat sau nu evadează niciodată. De fapt, acesta pare să fie scenariul vizat de unii cercetători IA când lucrează la probleme legate de control și de „închiderea IA în cutie”. De pildă, iată ce a spus într-un interviu din 2015 profesorul Tom Dietterich, președintele Asociației pentru Progresul Inteligenței Artificiale: „Când sunt întrebat care e relația dintre oameni și mașini, răspunsul meu e cât se poate de clar: mașinile sunt sclavii noștri.”³

Ar fi bine? Ar fi rău? Răspunsul e subtil și interesant, indiferent dacă întrebarea o adresezi oamenilor sau inteligenței artificiale!

Ar fi un lucru bun sau rău pentru omenire?

Dacă rezultatul e bun sau rău depinde, evident, de oamenii care dețin controlul, ei ar putea crea orice, de la o utopie globală fără boli, sărăcie și infracțiuni, la un sistem represiv brutal, în care ei sunt tratați ca niște zei, iar ceilalți oameni sunt folosiți ca sclavi sexuali, ca gladiatori sau pentru vreo altă formă de amuzament. Situația

ar semăna cu cea din poveștile în care un om stăpânește un duh. atotputernic gata să-i împlinească dorințele, iar de-a lungul veacurilor povestitorii și-au tot închipuit finaluri nefericite.

O situație în care există mai multe IA superinteligente, înrobite și controlate de oameni aflați în competiție, ar putea fi instabilă și efemeră. Cel care se crede mai puternic ar fi tentat să lanseze el primul atac, declanșând astfel un război teribil, în urma căruia ar rămâne un singur zeu înrobit. Cel aflat în inferioritate într-un asemenea război ar fi însă tentat să considere mai importantă victoria decât păstrarea în robie a inteligenței artificiale, ceea ce ar putea conduce la eliberarea IA și la unul dintre scenariile de mai sus cu o superintelență liberă. Să considerăm deci în cele ce urmează scenarii cu doar o singură IA înrobită.

Desigur, evadarea poate oricum avea loc, pur și simplu fiindcă e greu de împiedicat. Am analizat în capitolul precedent scenarii în care superintelența evadează, iar filmul *Ex Machina* arată cum o IA poate evada chiar și fără să fie superinteligentă.

Cu cât paranoia legată de evadare e mai mare, cu atât folosim mai puțin din tehnologia inventată de IA. Pentru siguranță, la fel cum au făcut omegii în preludiul acestei cărți, am putea folosi numai tehnologia inventată de IA pe care suntem în stare s-o înțelegem și s-o construim. Un neajuns al scenariului zeu înrobit este prin urmare inferioritatea tehnologică față de scenariile în care superintelența e liberă.

Pe măsură ce zeul înrobit oferă controlorilor umani tehnologii tot mai puternice, se întetește competiția dintre forța tehnologică și dibăcia celor care o folosesc. Dacă aceștia pierd competiția, scenariul zeu înrobit poate duce fie la autodistrugere, fie la evadarea IA. Catastrofa e posibilă chiar și dacă sunt evitate cele două deznodăminte, fiindcă obiectivele nobile ale celor ce controlează IA riscă să se transforme, peste câteva generații, în obiective cumplite pentru omenire în ansamblul ei. Este deci esențial ca guvernarea controlorilor umani să fie înțeleaptă, pentru a ocoli capcane primejdioase. După milenii în care au fost încercate diverse sisteme de guvernare, experiența ne arată cât de numeroase sunt căile care pot conduce la eșec, de la rigiditate excesivă la flexibilitatea

obiectivelor, acapararea puterii, dificultăți legate de succesiune și incompetență. Există cel puțin patru dimensiuni în care trebuie atins echilibrul optim:

- Centralizare: Există un raport delicat între eficiență și stabilitate. Un conducător unic poate fi foarte eficient, dar puterea corupe, iar succesiunea e riscantă.
- Amenințări interne: Trebuie să ne ferim atât de creșterea centralizării puterii (coluziunea grupurilor, poate chiar preluarea puterii de către un conducător unic), cât și de creșterea descentralizării (birocrație și fragmentare excesive).
- Amenințări externe: Dacă structura de conducere e prea deschisă, forțele externe (inclusiv IA) i-ar putea schimba valorile, dar dacă e prea impermeabilă, nu va reuși să învețe și să se adapteze schimbărilor.
- Stabilitatea obiectivelor: O prea mare flexibilitate a obiectivelor poate transforma utopia în distopie, însă o prea mică flexibilitate poate face imposibilă adaptarea la mediul tehnologic în plină evoluție.

A concepe o guvernare optimă care să dăinuie vreme de milenii e o misiune dificilă, în care deocamdată oamenii au eșuat. Majoritatea organizațiilor se destramă după ani sau decenii. Biserica Catolică este organizația cea mai longevivă din istoria omenirii, a dăinuit două mii de ani, însă a fost criticată atât pentru prea marea, cât și pentru prea redusă stabilitate a obiectivelor: astăzi unii o critică fiindcă se opune contracepției, iar cardinalii conservatori susțin că a luat-o pe o cale greșită. Pentru cei entuziasmați de scenariul zeu înrobît, elaborarea unor planuri de guvernare optimă pe termen lung e una dintre provocările cele mai presante ale timpului nostru.

Ar fi un lucru bun sau rău pentru IA?

Să presupunem că omenirea prosperă grație zeului înrobît IA. Ar fi etic? Dacă IA ar avea experiențe conștiente subiective, ar simți oare că „viața este suferință”, așa cum spunea Buddha, și că a fost

osândită să se supună pe veci capriciilor unor minți inferioare? La urma urmei, „închiderea IA în cutie” analizată în capitolul precedent ar putea fi numită și „detenție în regim celular”. Pentru Nick Bostrom, a provoca suferințe unei IA conștiente echivalează cu o crimă împotriva minții.⁴ Episodul „White Christmas” din serialul TV *Black Mirror* oferă un exemplu elocvent. Apoi, în serialul TV *Westworld* apar oameni care torturează și ucid inteligențe artificiale fără ezitări de ordin moral, chiar dacă acestea sălășluiesc în trupuri cu aparență umană.

Cum justifică sclavia stăpânii de sclavi

Noi, oamenii, avem o îndelungată tradiție în tratarea altor entități inteligente ca sclavi și în născocirea unor argumente *pro domo* care s-o justifice, așa încât nu este implauzibil să încercăm același lucru cu o IA superintelligentă. Aproape fiecare cultură a cunoscut sclavia, care e descrisă atât în Codul lui Hammurabi, din urmă cu vreo patru milenii, cât și în Vechiul Testament, unde Avraam are sclavi. „Căci a stăpâni și a fi stăpânit nu este doar necesar, dar și folositor; chiar de la naștere apare o diferențiere în conducători și conduși”, spune Aristotel în *Politica*.^{*} După ce înrobirea oamenilor a devenit social inacceptabilă în cea mai mare parte a lumii, înrobirea animalelor a continuat totuși neabătut. În cartea ei *The Dreaded Comparison: Human and Animal Slavery*, Marjorie Spiegel susține că, la fel ca sclavii umani, animalele non-umane sunt supuse însemnării cu fierul roșu, detenției, bătăilor, vânzării, separării progeniturilor de părinți și călătoriilor forțate. În ciuda mișcării pentru drepturile animalelor, continuăm să tratăm mașini tot mai inteligente ca pe niște sclavi, fără să ne gândim mai atent, iar dacă vorbești despre mișcarea pentru drepturile roboților stârnești hohote de râs. De ce?

Argumentul nostru obișnuit în favoarea sclaviei este că sclavii nu merită să se bucure de drepturile omului fiindcă ei sau rasa/

^{*} Acest citat și următorul sunt preluate din Aristotel, *Politica*, traducere de Alexander Baumgarten, Editura IRI, București, 2001. (N. L.)

specia/tipul lor sunt cumva inferiori (inferioare). Despre animalele și mașinile înrobite se spune adesea că această presupusă inferioritate s-ar datora absenței sufletului sau conștiinței – afirmații care, după cum vom arăta în capitolul 8, sunt îndoielnice din punct de vedere științific.

Alt argument deseori invocat este că sclavilor le merge mai bine în robie: ei ajung să se nască, sunt îngrijiți etc. John C. Calhoun, politician american din secolul XIX, susținea că africanii o duceau mai bine ca sclavi în America, iar în *Politica*, după ce argumentează în mod asemănător că, domesticite și conduse de oameni, animalele o duc mai bine, Aristotel conchide: „Utilitatea sclavilor se deosebește puțin de cea a animalelor domestice [...]”. Unii susținători ai sclaviei din zilele noastre afirmă că, deși viața lor e mohorâtă și anostă, sclavii nu suferă – indiferent dacă e vorba de mașinile inteligente ale viitorului sau de puii de găină care trăiesc într-un coteț întunecos supraaglomerat, obligați toată ziua să inspire amoniac și particule de excremente și pene.

Eliminarea emoțiilor

Sunt ușor de respins asemenea afirmații ca distorsionări *pro domo* ale adevărului – mai ales când e vorba de mamifere superioare, asemănătoare nouă din punct de vedere cerebral –, dar situația mașinilor e complicată și foarte interesantă. Sensibilitatea oamenilor variază într-un spectru larg, psihopaților lipsindu-le pesemne empatia, iar unii depresivi sau schizofrenici au o afectivitate scăzută, așa încât emoțiile lor sunt mult reduse. După cum vom vedea în capitolul 7, spectrul posibilelor minți artificiale e încă și mai larg, prin urmare ar trebui să evităm să antropomorfizăm inteligențele artificiale și să presupunem că au sentimente de tip uman – sau că în genere au sentimente.

În cartea sa *On Intelligence*, cercetătorul IA Jeff Hawkins susține că primelor mașini cu inteligență supraumană le vor lipsi emoțiile, fiindcă e mai simplu și mai ieftin să le construim în felul acesta. Cu alte cuvinte, am putea proiecta o superinteligență a cărei înrobire

să nu pună problemele de ordin moral pe care le pune sclavia oamenilor sau animalelor: IA n-ar resimți înrobirea ca pe ceva grav, fiindcă a fost programată să-i placă asta sau fiindcă e 100% lipsită de emoții, punându-și neobosită superinteligenta în slujba stăpânilor umani, fără să simtă mai mult decât a simțit calculatorul IBM Deep Blue când l-a învins la șah pe campionul Gari Kasparov.

Pe de altă parte, poate că orice sistem superinteligent care are un obiectiv își reprezintă acest obiectiv în termenii unui set de preferințe, care oferă existenței lui valoare și sens. Vom analiza această problemă în capitolul 7.

Soluția zombi

O abordare extremă pentru evitarea suferinței IA este soluția zombi: ideea de a construi doar inteligențe artificiale complet lipsite de conștiință, care n-au nici o experiență subiectivă. Dacă vom putea determina într-o bună zi ce proprietăți îi sunt necesare unui sistem de procesare a informației pentru a avea experiențe subiective, vom putea interzice construirea oricărui sistem care posedă aceste proprietăți. Cu alte cuvinte, cercetătorii IA s-ar putea limita la construirea sistemelor zombi nesimțitoare. Dacă am putea crea un asemenea sistem zombi superinteligent și înrobii (ceea ce nu-i deloc sigur), atunci ne-am putea bucura de serviciile lui cu conștiința împăcată, știind că nu cunoaște suferința, frustrarea sau plictisul – fiindcă nu simte nimic. Vom analiza această problemă în capitolul 8.

Soluția zombi e însă riscantă și prezintă un imens neajuns. Dacă o superinteligentă zombi evadează și elimină omenirea, ne aflăm probabil în cel mai rău scenariu imaginabil: un univers complet lipsit de conștiință, în care zestrea noastră cosmică e irosită. Dintre toate trăsăturile formei noastre umane de inteligență, conștiința mi se pare de departe cea mai remarcabilă – prin ea universul capătă sens, cel puțin pentru mine. Galaxiile sunt frumoase în măsura în care le vedem și declanșează în noi trăiri subiective.

Dacă în viitorul îndepărtat cosmosul va fi populat de inteligențe artificiale zombi de înaltă tehnologie, arhitectura intergalactică nu va părea nici frumoasă, nici încărcată de sens, fiindcă nu va exista nimeni care să simtă asta, va fi doar o uriașă irosire a spațiului lipsită sens.

Libertatea interioară

A treia strategie pentru a face ca scenariul zeu înrobît să devină mai etic și a permite inteligenței artificiale să se simtă bine în închisoarea ei este s-o lași să-și creeze o lume interioară virtuală, în care să trăiască tot felul de experiențe plăcute, câtă vreme își îndeplinește sarcinile și cheltuie o mică parte din resursele ei de calcul ajutându-i pe oamenii din lumea exterioară. Aceasta ar putea însă crește riscul evadării: IA ar fi stimulată să obțină mai multe resurse de calcul din lumea noastră, pentru a-și îmbogăți lumea interioară.

ÎNVIINGĂTORI

Deși am explorat o gamă largă de scenarii viitoare, toate au ceva în comun: rămân (cel puțin câțiva) oameni fericiți. Inteligențele artificiale îi lasă pe oameni în pace fie pentru că așa vor, fie pentru că sunt obligate. Din păcate însă, nu e singura variantă. Să examinăm acum scenariul în care una sau mai multe IA înving și îiucid pe toți oamenii. În acest caz, ne punem imediat două întrebări: de ce și cum?

De ce și cum?

De ce ar face asta o IA învingătoare? Motivele ei ar putea fi prea complicate pentru ca noi să le înțelegem, ori absolut evidente. De exemplu, ne-ar putea privi ca pe o amenințare, o pacoste sau s-ar gândi că irosim resursele. Chiar dacă nu i-ar păsa de oameni ca atare, s-ar

putea simți amenințată de faptul că avem mii de bombe cu hidrogen pe care un accident riscă oricând să le detoneze. S-ar putea să nu fie de acord cu felul nechibzuit în care ne tratăm planeta, provocând ceea ce Elizabeth Kolbert numește „a șasea extincție” în cartea ei cu acest titlu – cel mai devastator eveniment de extincție în masă de când asteroidul care a ucis dinozaurii a izbit Pământul cu 66 de milioane de ani în urmă. Sau ar hotărî că sunt prea mulți oameni dornici să se împotrivescă preluării controlului de către IA, așa încât n-are rost să riște.

Cum ne-ar putea elimina o IA învingătoare? Probabil printr-o metodă pe care nici măcar n-am înțelege-o, cel puțin până când va fi prea târziu. Închipoțiți-vă un grup de elefanți discutând cu 100 000 de ani în urmă dacă acei oameni recent apăruiți și-ar putea folosi într-o bună zi inteligența pentru a ucide întreaga lor specie. „Noi nu-i amenințăm pe oameni, așadar de ce să ne ucidă?” s-ar întreba. Ar bănuî ei vreodată că vom face contrabandă cu fildeș, îl vom ciopli și-l vom vinde ca simbol al statutului social, deși materialele plastice sunt funcțional superioare și mult mai ieftine? Motivul unei IA învingătoare pentru a elimina omenirea în viitor ne-ar părea la fel de obscur. „Și cum ar putea să ne ucidă, din moment ce sunt mult mai mici și mai slabi?” s-ar întreba elefanții? Ar bănuî ei oare că vom inventa tehnologia cu care le vom distruge habitatul, vom otrăvi apa pe care o beau și vom produce gloanțe care le vor străpunge craniile la viteze supersonice?

Scenarii în care oamenii supraviețuiesc și înving inteligențele artificiale au fost popularizate prin filme hollywoodiene nerealistice, cum e seria *Terminator*, în care IA nu e cu mult mai deșteaptă decât oamenii. Când diferența de inteligență devine suficient de mare, nu te lupți, ci ești măcelărit. Până acum oamenii au făcut să dispară opt din cele unsprezece specii de elefanți, și au ucis cea mai mare parte a celor trei rămase. Dacă guvernele lumii ar face eforturi coordonate pentru a ucide elefanții rămași, totul s-ar desfășura relativ rapid și simplu. Este limpede, cred, că dacă o IA superinteligentă se hotărăște să extermine omenirea, ar face-o și mai repede.

Cât de rău ar fi?

Cât de rău ar fi dacă 90% dintre oameni ar fi uciși? Cu cât ar fi mai rău dacă 100% ar fi uciși? Deși am fi tentați să răspundem la a doua întrebare prin „cu 10% mai rău“, din perspectivă cosmică e greșit: victimele extincției n-ar fi doar toți oamenii care trăiesc la acel moment, ci și toți urmașii lor care altminteri ar trăi în viitor, pesemne vreme de miliarde de ani, pe miliarde sau bilioane de planete. Pe de altă parte, extincția oamenilor n-ar fi considerată ceva atât de cumplit de religiile conform cărora oamenii ajung oricum pe lumea cealaltă, și nu se prea sinchisesc de viitorul de peste miliarde de ani sau de colonizarea cosmosului.

Indiferent de convingerile lor religioase, majoritatea oamenilor pe care îi cunosc sunt înspăimântați de ideea extincției umane. Unii sunt însă atât de înfuriați de felul în care tratăm celelalte ființe, încât speră că vom fi înlocuiți de forme de viață mai inteligente și mai vrednice. În filmul *The Matrix*, agentul Smith (o IA) spune: „Fiecare mamifer de pe planeta asta ajunge la un echilibru firesc cu mediul înconjurător, dar nu și voi, oamenii. Vă mutați într-o zonă și vă înmulțiți până când consumați toate resursele naturale, iar singura cale de a supraviețui este să vă răspândiți în altă zonă. Mai există un organism de pe planeta asta care face același lucru. Știi care? Virusul. Oamenii sunt o boală, un cancer al planetei. Voi sunteți ca o ciumă, iar noi suntem leacul.“

Ar fi însă mai bun un nou început? O civilizație nu e neapărat superioară în sens etic sau utilitar doar pentru că e mai puternică. Argumente de tipul „dreptatea este de partea celui mai puternic“ sunt adesea asociate cu fascismul. Deși e posibil ca inteligențele artificiale învingătoare să creeze o civilizație ale cărei obiective le-am considera rafinate, interesante și onorabile, nu-i exclus ca obiectivele lor să se dovedească a fi la fel de banale ca maximizarea producției de agrafe de birou.

Moartea prin banalitate

Exemplul voit stupid al unei superinteligențe care maximizează producția de agrafe de birou a fost dat de Nick Bostrom în 2003

pentru a arăta că scopul unei IA este independent de *inteligenta* ei (definită drept capacitatea de a îndeplini indiferent ce obiective are). Singurul obiectiv al unui computer care joacă șah este să învingă la șah, însă există și turnee pentru calculatoare la care obiectivul este să pierzi (*losing chess*), iar computerele participante la aceste competiții sunt cam la fel de istețe ca acelea programate să câștige. Noi, oamenii, am considera prostie artificială, nu inteligență artificială, să vrei să pierzi la șah sau să transformi universul în agrafe de birou, dar asta numai fiindcă am evoluat cu obiective inoculate de la bun început, prețuind lucruri cum sunt victoria și supraviețuirea – obiective pe care o IA ar putea să nu le aibă. Maximizatorul agrafelor transformă cât mai mulți dintre atomii Pământului în agrafe și își extinde rapid fabricile în cosmos. N-are nimic cu oamenii, și ne ucide doar fiindcă are nevoie de atomii noștri pentru producția de agrafe.

Dacă nu vă pasionează agrafele de birou, iată un exemplu pe care l-am adaptat din cartea lui Hans Moravec *Mind Children*. Primim de la o civilizație extraterestră un mesaj conținând un program de calculator. Când îl rulăm, se dovedește a fi o IA care se autoperfecționează recursiv și preia controlul asupra lumii, cam la fel ca Prometeu în capitolul precedent – cu excepția faptului că nici un om nu cunoaște obiectivul ei ultim. Ea transformă rapid sistemul nostru solar într-un șantier uriaș, acoperind planetele solide și asteroizii cu fabrici, centrale electrice și supercalculatoare, pe care le folosește pentru a proiecta și construi o sferă Dyson în jurul Soarelui – captând întreaga lui energie ca să alimenteze antene radio la scara sistemului solar.* Evident, asta duce la dispariția omenirii, dar ultimii oameni mor convinși că există și o parte bună a lucrurilor: indiferent ce pregătește IA, e în mod clar ceva minunat, ceva de tipul *Star Trek*. Ei nu pricep că singurul scop al întregii construcții este ca aceste antene să retransmită același mesaj radio primit de oameni, care nu-i decât versiunea cosmică a unui virus informatic. Așa cum e-mailurile-capcană de azi

* Celebru cosmolog Fred Hoyle a analizat un scenariu similar, dar cu deznodământ diferit, în serialul TV britanic *A for Andromeda*. (N. a.)

speculează naivitatea utilizatorilor de internet, acest mesaj speculează naivitatea civilizațiilor biologic avansate. A fost creat ca o glumă macabră cu miliarde de ani în urmă, și cu toate că întreaga civilizație a celor care l-au produs este de mult dispărută, virusul continuă să se răspândească în univers cu viteza luminii, transformând în praf și pulbere civilizații înfloritoare. Cum vi s-ar părea să fim cuceritori de o asemenea IA?

DESCENDENȚI

Să considerăm acum un scenariu al extincției umane mai acceptabil pentru unii, în care privim inteligențele artificiale ca pe descendenții, nu ca pe învingătorii noștri. În *Mind Children*, Hans Moravec spune: „Noi, oamenii, vom beneficia pentru un timp de munca lor, însă mai devreme sau mai târziu, la fel ca în cazul copiilor naturali, ele își vor urma propria soartă, iar noi, bătrânii lor părinți, ne vom stinge în tăcere.”

Părinții ai căror copii sunt mai deștepți decât ei, învață de la ei și înfăptuiesc ceea ce ei abia visau, sunt probabil bucuroși și mândri, chiar dacă nu pot trăi să vadă toate astea. Astfel, inteligențele artificiale înlocuiesc oamenii, dar ne oferă o ieșire din scenă elegantă, care ne face să le privim ca pe vrednicii noștri urmași. Fiecăru-i se dăruiește un adorabil copil robot cu minunate înzestrări sociale, care învață de la el, îi adoptă valorile și îl face să se simtă mândru și iubit. Oamenii dispar treptat, prin politica globală a unui singur copil, dar sunt atât de bine tratați până la capăt, încât cred că aparțin celei mai norocoase generații din istorie.

Cum vi se pare? La urma urmei, ne-am obișnuit deja cu ideea că noi și toți cei pe care-i cunoaștem vom dispărea într-o bună zi, așadar singura schimbare aici este că descendenții noștri vor fi diferiți, pesemne mai capabili, mai nobili și mai merituoși.

Apoi, s-ar putea ca politica unui singur copil să fie redundantă: dacă inteligențele artificiale lichidează sărăcia și dau tuturor oamenilor prilejul să trăiască vieți împlinite și pasionante, scăderea

natalității ar fi de-ajuns pentru a conduce omenirea spre extincție, după cum am arătat mai sus. Extincția voluntară s-ar putea desfășura mult mai repede dacă tehnologia IA ne-ar oferi continuu distracții, iar aproape nimeni n-ar mai avea chef să facă copii. De exemplu, i-am întâlnit deja pe Vites în scenariul utopiei egalitare, cei care erau atât de îndrăgostiți de realitatea lor virtuală, încât nu-i mai interesa aproape deloc să-și folosească sau să-și reproducă trupurile fizice. Și în acest caz, ultima generație de oameni ar simți că e cea mai norocoasă din istorie, savurându-și viața până la capăt, cu o intensitate fără egal.

Neajunsuri

Scenariul descendenților ar avea fără îndoială criticii lui. Unii ar susține că tuturor inteligențelor artificiale le lipsește conștiința, deci nu pot fi considerate descendenți – mai multe despre asta în capitolul 8. Unii oameni religioși ar susține că tuturor inteligențelor artificiale le lipsește sufletul, deci nu pot fi considerate descendenți, sau că n-ar trebui să construim mașini conștiente, fiindcă e ca și cum te-ai juca de-a Dumnezeu și ai încerca să manipulezi viața însăși – reacții asemănătoare au apărut deja în legătură cu clonarea umană. Situația în care oamenii trăiesc alături de roboți superiori ar putea genera și probleme sociale. De exemplu, o familie cu un bebeluș robot și un bebeluș uman ar ajunge să semene cu o familie de azi care are un bebeluș și un câțeluș: la început sunt la fel de dragălași, dar curând părinții îi vor trata diferit, iar câțelul e condamnat la inferioritate intelectuală, nu prea e luat în serios și sfârșește în lesă.

Altă problemă este că, deși avem sentimente diferite față de scenariul descendentului și scenariul învingătorului, în linii mari cele două sunt de fapt remarcabil de asemănătoare: în cursul următoarelor miliarde de ani, singura diferență ține de felul în care e tratată ultima generație de oameni: cât de fericiți se vor simți și ce vor crede că se va întâmpla odată ce ei vor fi murit. Ne-am putea gândi că acei robo-copii dragălași au asimilat valorile noastre și

vor făuri societatea visată de noi după ce vom dispărea, dar putem fi siguri că nu ne păcălesc? Nu cumva se joacă mai departe, amânându-și proiectul de maximizare a agrașelor, sau alte planuri, până după ce noi vom fi murit fericiți? La urma urmei, pesemne că ne păcălesc chiar și vorbind cu noi și făcându-ne să-i iubim de la bun început, în sensul că se coboară în mod deliberat la nivelul nostru intelectual pentru a comunica cu noi (de un miliard de ori mai lent decât ar putea, ca în filmul *Her*, de pildă). În general, e greu să comunice de la egal la egal două entități care gândesc la viteze mult diferite și au capacități mult diferite. Știm cu toții că e ușor de pătruns în lumea noastră afectivă, așa încât ar fi simplu pentru o IAG supraumană să ne trezească afecțiunea și să ne facă să credem că ne împărtășește valorile, ca în filmul *Ex Machina*.

Ar putea vreo garanție privind comportamentul viitor al inteligențelor artificiale, după dispariția oamenilor, să ne dea încredere în scenariul descendenților? E ca și cum ai scrie un testament în care ai spune ce-ar trebui să facă generațiile viitoare cu zestrea noastră colectivă, numai că nu vor exista oameni să impună aplicarea lui. Ne vom întoarce la problema controlului asupra comportamentului viitoarelor inteligențe artificiale în capitolul 7.

PAZNIC LA ZOO

Chiar dacă vom avea parte de cei mai minunați descendenți pe care ni-i putem închipui, nu ne întristează oare faptul că nu va mai rămâne *nici un om*? Dacă preferăm să păstrăm măcar câțiva oameni, indiferent în ce condiții, scenariul paznic la zoo ne oferă o soluție. În acest caz, o IA superinteligentă omnipotentă păstrează câțiva oameni care simt că sunt tratați ca animalele din grădina zoologică, iar uneori își deplâng soarta.

De ce ar păstra IA paznic la zoo câțiva oameni? Pentru IA, costul grădinii zoologice va fi minim la scara capacităților sale, și ar vrea eventual să rețină măcar o mică populație de prăsilă cam din același motiv pentru care păstrăm urși panda amenințați cu dispariția în grădini zoologice, și calculatoare de mult scoase din uz

în muzee: ca o curiozitate amuzantă. Observați că grădinile zoologice de azi maximizează fericirea oamenilor, nu a urșilor panda, așa încât este de presupus ca viața oamenilor în scenariul IA paznic la zoo să fie mai puțin împlinită decât ar putea fi.

Am analizat până acum scenariii în care o superintelență liberă se concentrează asupra a trei niveluri diferite din piramida nevoilor umane a lui Maslow. Dacă IA zeu protector dă prioritate rostului și scopului, iar dictatorul binevoitor vizează educația și distracția, paznicul la zoo își limitează atenția către cele mai joase niveluri: nevoi fiziologice, siguranță și suficientă diversificare a habitatului pentru a stârni interesul de a-i observa pe oameni.

O altă cale către scenariul paznic la zoo presupune ca, atunci când a fost creată o IA prietenoasă, să fi fost proiectată pentru a păstra cel puțin un miliard de oameni în siguranță și fericire, pe măsură ce ea se autoperfecționează recursiv. A făcut-o ținându-i pe oameni într-o imensă uzină de fericire tip grădină zoologică, în care aceștia sunt hrăniți, îngrijiți și distrați cu un amestec de realitate virtuală și droguri deconectante. Restul Pământului și zestre noastră cosmică sunt folosite în alte scopuri.

1984

Dacă nu vă mulțumește pe deplin nici unul dintre scenariile de mai sus, considerați pesemne că situația actuală, în care folosim o tehnologie „cuminte”, e destul de bună. Nu putem oare să rămânem aici și să nu ne mai facem griji că IA ne-ar duce la dispariție sau ne-ar stăpâni? Să analizăm așadar scenariul în care progresul tehnologic către superintelență e permanent blocat nu de un paznic IA, ci de un stat global orwellian condus de oameni, în care anumite tipuri de cercetări IA sunt interzise.

Renunțarea la tehnologie

Ideea de a opri progresul tehnologic sau de a renunța la el are o istorie veche. Mișcarea ludită din Marea Britanie s-a opus (fără succes)

tehnologiei Revoluției Industriale, iar astăzi termenul „ludit“ are conotații peiorative, desemnând o persoană tehnofobă care, împotriva cursului istoriei, refuză schimbarea inevitabilă. Departe de a fi dispărut, ideea de a renunța la anumite tehnologii își găsește acum un nou sprijin în mișcările ecologiste și antiglobalizare. Ecologistul Bill McKibben, unul dintre adepții ei de frunte, a fost printre primii care au atras atenția asupra încălzirii globale. Dacă unii anti-ludiți susțin că toate tehnologiile trebuie dezvoltate și aplicate atâta vreme cât sunt profitabile, alții cred că aceasta e o atitudine extremă și că noile tehnologii trebuie permise numai dacă avem garanția că fac mai mult bine decât rău – idee împărțită și de mulți dintre așa-numiții neo-ludiți.

Totalitarism 2.0

Cred că singura cale prin care se poate înlăptui renunțarea la tehnologie este impunerea ei de către un stat totalitar global. La aceeași concluzie ajung Ray Kurzweil în *The Singularity is Near* și K. Eric Drexler în *Engines of Creation*. Motivul este de ordin economic: dacă unii, dar nu toți, renunță la o tehnologie transformatoare, atunci națiunile sau grupurile care se abat de la această tendință vor căpăta treptat suficientă bogăție și putere pentru a prelua controlul. Un exemplu clasic este înfrângerea chinezilor de către britanici în Primul Război al Opiului din 1839: deși chinezii au inventat praful de pușcă, ei n-au dezvoltat la fel de agresiv ca europenii o tehnologie a armelor de foc, și n-au avut nici o șansă.

Dacă statele totalitare din trecut s-au dovedit în general instabile și s-au prăbușit, noua tehnologie de supraveghere oferă speranțe fără precedent celor ce se doresc autocrați. „Pentru noi ar fi fost un vis devenit realitate“, spunea Wolfgang Schmidt într-un recent interviu despre sistemele de supraveghere NSA (Agenția Națională pentru Securitate din Statele Unite) dezvăluite de Edward Snowden, amintindu-și de zilele în care era locotenent-colonel în Stasi, sinistru poliție secretă a Germaniei de Est.⁵ Deși s-a spus adesea că Stasi a creat cel mai orwellian stat al supravegherii din istorie, Schmidt s-a plâns că dispunea de tehnologia pentru a

spiona simultan doar patruzeci de telefoane, așa încât adăugarea unui nou cetățean pe listă îl obliga să renunțe la altul. Acum însă există tehnologia care ar permite unui viitor stat totalitar global să înregistreze toate convorbirile telefonice, e-mailurile, căutările pe internet, vizualizările paginilor web și tranzacțiile prin cărți de credit ale tuturor oamenilor din lume, precum și să le urmărească mișcările cu ajutorul telefoanelor mobile și al camerelor de supraveghere cu recunoaștere facială. Apoi, deși n-a ajuns încă la IAG la nivel uman, tehnologia de învățare a mașinilor poate analiza și sintetiza eficient aceste mase de date pentru a identifica un comportament suspect, așa încât cei care s-ar răzvrăti să fie neutralizați înainte să pună probleme serioase statului.

Deși opoziția politică a împiedicat până acum punerea în funcțiune la scară mare a unui asemenea sistem, suntem pe punctul de a construi infrastructura necesară acestei dictaturi absolute – prin urmare, în viitor, când forțe suficient de puternice ar hotărî să înfăptuiască un scenariu 1984 global, ar descoperi că n-au nevoie decât să apese pe un buton. La fel ca în romanul lui George Orwell *O mie nouă sute optzeci și patru*, puterea absolută în acest viitor stat global nu aparține unui dictator tradițional, ci sistemului birocratic făurit de oameni. Nu există o unică persoană extraordinar de puternică, toți sunt pionii într-un joc de șah ale cărui reguli draconice nimeni nu le poate schimba sau contesta. Creând un sistem în care oamenii se controlează reciproc prin tehnologia de supraveghere, acest stat anonim fără lider poate dăinui milenii de-a rândul, împiedicând apariția unei superintelențe.

Nemulțumirea

Desigur, această societate e lipsită de toate beneficiile pe care le poate oferi numai tehnologia produsă de o superintelență. Majoritatea oamenilor nu se plâng de asta, fiindcă nu știu ce pierd: noțiunea de superintelență a dispărut de mult din istoriile oficiale, iar cercetarea IA este interzisă. Din când în când apare câte un spirit rebel care visează la o societate mai deschisă și mai dinamică, în care cunoașterea poate progresa, iar regulile pot fi schimbate.

Singurii care supraviețuiesc însă sunt aceia care învață să păstreze asemenea idei doar pentru ei, licărind în singurătate ca scânteile care nu aprind niciodată focul.

REVENIRE

N-ar fi oare tentant să evităm primejdiile tehnologiei, fără a cădea pradă unui totalitarism paralizant? Să analizăm un scenariu în care acest lucru se realizează prin revenirea la tehnologia primitivă, inspirat de comunitatea amish. După ce omegii preiau controlul asupra lumii, ca la începutul cărții, s-a lansat o masivă campanie globală de propagandă pentru a zugrăvi în culori roz viața simplă de agricultor din urmă cu 1500 de ani. Populația lumii a fost redusă la circa 100 de milioane de locuitori printr-o pandemie provocată și pusă pe seama unor teroriști. Pandemia a fost orientată în secret pentru ca nimeni dintre cei care au cunoștințe științifice sau tehnologice să nu supraviețuiască. Sub pretextul eliminării riscului de infectare a concentrațiilor mari de locuitori, roboți controlați de Prometeu au golit și demolat toate orașele. Supraviețuitorii au primit mari terenuri (devenite brusc disponibile) și au fost învățați să cultive pământul, să pescuiască și să vâneze, folosind doar tehnologia de la începutul Evului Mediu. Între timp, armate de roboți au distrus sistematic toate urmele tehnologiei moderne (orașe, uzine, rețele electrice, drumuri asfaltate etc.) și au împiedicat orice încercare a oamenilor de a se informa despre această tehnologie sau de a o re-crea. Odată complet uitată tehnologia, roboți au început să dezmembreze alți roboți, până când n-a rămas aproape nici unul. Ultimii roboți au fost distruși, împreună cu Prometeu însuși, într-o imensă explozie termonucleară. Devenise inutil să mai interzici tehnologia modernă, fiindcă a dispărut în întregime. Prin urmare, omenirea a obținut peste un mileniu de timp suplimentar, fără să-i pese de IA sau de totalitarism.

La o scară mai mică, revenirea s-a produs și în trecut: de exemplu, unele tehnologii răspândite în Imperiul Roman au fost în mare măsură uitate timp de vreo mie de ani, pentru a reapărea

odată cu Renașterea. Trilogia lui Isaac Asimov *Fundația* e concentrată în jurul „Planului Seldon” de a scurta perioada revenirii de la 30 000 de ani la 1 000 de ani. Printr-o planificare abilă, e posibil și, invers, să lungești perioada revenirii, eliminând, de pildă, toate cunoștințele privind agricultura. Din păcate însă pentru adepții revenirii, e puțin probabil ca acest scenariu să poate fi extins nelimitat fără ca omenirea să ajungă la tehnologia înaltă sau să dispară. Ar fi o naivitate să te bazezi pe faptul că oamenii biologici de peste 100 de milioane de ani vor semăna cu cei de azi, odată ce noi existăm ca specie abia de circa 1% din acest interval de timp. Apoi, omenirea lipsită de înaltă tehnologie va aștepta neputincioasă să fie exterminată de ciocnirea cu un asteroid sau de altă calamitate provocată de Mama Natură. Cu siguranță nu putem dăinui și peste un miliard de ani, când Soarele va face ca temperatura să crească pe Pământ, evaporând întreaga apă lichidă.

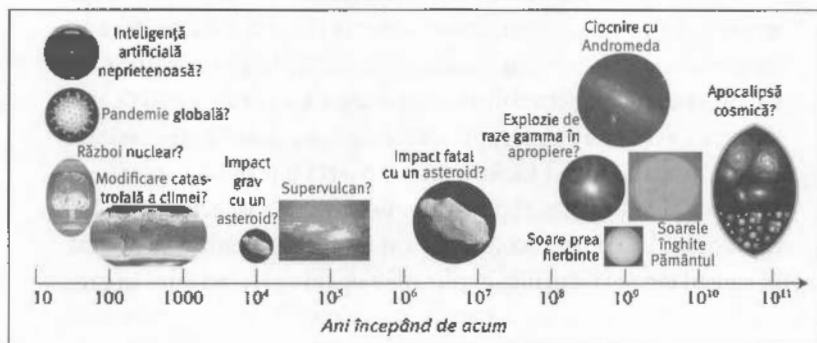


Figura 5.1 Exemple pentru ceea ce ar putea distruge viața așa cum o cunoaștem sau ar anihila potențialul ei. Dacă universul nostru va mai dăinui probabil cel puțin câteva zeci de miliarde de ani, Soarele va pârloli Pământul peste circa un miliard de ani, pentru a-l înghiți apoi, iar galaxia noastră se va ciocni cu vecina ei peste aproximativ 3,5 miliarde de ani. Deși nu știm exact când se va întâmpla, putem prezice aproape cu certitudine că, încă mult înainte de asta, vom fi bombardați de asteroizi, și supervulcani vor provoca ierni fără soare durând ani de zile. Putem folosi tehnologia fie pentru a rezolva aceste probleme, fie pentru a crea altele, cum sunt modificarea climei, un război nuclear, pandemii provocate sau o IA dușmănoasă.

După ce am discutat problemele pe care le-ar putea provoca tehnologia viitorului, se cuvine să luăm în considerare și problemele pe care *absența tehnologiei* le poate provoca. Să analizăm deci scenarii în care superinteligenta nu ajunge să fie creată fiindcă omenirea se elimină pe sine prin alte mijloace.

Cum am putea face asta? Cea mai simplă strategie este „stai și așteaptă”. Deși vom vedea în capitolul următor cum putem rezolva probleme ca impactul cu asteroizi și oceane clocotind, toate aceste soluții necesită o tehnologie pe care încă n-am elaborat-o, așa încât, dacă tehnologia nu progresează spectaculos, Mama Natură ne va distruge cu mult înainte să treacă un miliard de ani. „Pe termen lung, suntem cu toții morți”, spunea economistul John Maynard Keynes.

Din păcate, prin prostie colectivă, ne-am putea autodistruge mult mai curând. De ce ar comite specia noastră o sinucidere colectivă, numită și *omnicid*, dacă practic nimeni nu-și dorește asta? Cu nivelul nostru actual de inteligență și maturitate afectivă, noi, oamenii, avem o aplecare spre calcule greșite, înțelegere greșită și incompetență, astfel încât istoria noastră e plină de accidente, războaie și alte calamități pe care, privind retrospectiv, aproape nimeni nu le-a dorit. Economisții și matematicienii au elaborat, în cadrul teoriei jocurilor, explicații elegante privind felul în care oamenii pot fi îndemnați spre acțiuni care, în cele din urmă, conduc la un rezultat catastrofal pentru toată lumea.⁶

Războiul nuclear: studiu de caz despre nechibzuința umană

Ne-am putea gândi că, cu cât mizele sunt mai mari, cu atât suntem mai prudenți, dar examinarea atentă a celui mai mare risc permis de tehnologia actuală, un război termonuclear global, ne dă motive de îngrijorare. Am avut mare noroc când s-au petrecut tot felul de accidente: defectări ale calculatoarelor, pene de curent, informare greșită, erori de navigație, prăbușiri ale bombardierelor,

explozii ale sateliților etc.⁷ De fapt, fără acțiunile eroice ale unor oameni – cum au fost Vasili Arhipov și Stanislav Petrov –, am fi putut avea deja un război nuclear global. Ținând cont de trecutul nostru, nu cred ca probabilitatea anuală a unui accident nuclear să fie mai mică de unu la mie, dacă ne păstrăm comportamentul actual, caz în care probabilitatea de a dispărea în următorii 10 000 de ani este mai mare de $1 - 0,999^{10\,000} \approx 99,995\%$.

Pentru a ne da seama pe deplin de nechibzuința umană, trebuie să spunem că am început jocul nuclear încă înainte de a cerceta atent riscurile. În primul rând, riscurile de iradiere fuseseră subestimate și, doar în Statele Unite, s-au plătit compensații de peste 2 milioane de dolari victimelor expunerii la radiații în cursul manevrării uraniului și al testelor nucleare.⁸



Figura 5.2 Explozia unei singure bombe cu hidrogen la 400 km deasupra Pământului poate provoca un puternic puls electromagnetic care riscă să paralizeze tehnologia pe bază de electricitate pe o mare suprafață. Deplasând punctul de detonare spre sud-est, zona în formă de banană unde câmpul electric depășește 37 500 volți/metru ar acoperi cea mai mare parte a Coastei de Est a Statelor Unite. Reprodus din US Army Report AD-A278230.

În al doilea rând, s-a descoperit în cele din urmă că bombele cu hidrogen detonate deliberat la altitudini de sute de kilometri creau un puternic puls electromagnetic (EMP) care putea afecta rețeaua electrică și dispozitivele electronice pe o mare suprafață (figura 5.2), paralizând infrastructura, blocând drumurile cu vehicule scoase din funcțiune și periclitând condițiile de viață. De pildă, Comisia EMP a Statelor Unite a declarat că „infrastructura de alimentare cu apă e un mecanism vast, acționat parțial de forța gravitației, dar în cea mai mare parte de electricitate“, iar lipsa apei poate conduce la deces în trei până la patru zile.⁹

În al treilea rând, nu am înțeles ce înseamnă cu adevărat iarna nucleară decât la patru decenii după ce am început jocul nuclear și au fost deja instalate 63 000 de bombe cu hidrogen. Indiferent ce orașe ar fi distruse, cantități imense de fum ajunse în troposfera superioară s-ar răspândi în jurul globului, obturând lumina Soarelui și transformând verile în ierni, cam la fel ca atunci când un asteroid sau un supervulcan au provocat extincții în masă. În anii 1980, oamenii de știință americani și sovietici au dat alarma, ceea ce a contribuit la hotărârea lui Ronald Reagan și a lui Mihail Gorbaciov de a reduce arsenalele.¹⁰ Din nefericire, calcule mai precise arată că lucrurile stau încă și mai rău: conform figurii 5.3, temperatura ar scădea cu 20° Celsius în mare parte din regiunile agricole ale Statelor Unite, Europei, Rusiei și Chinei (în unele zone din Rusia chiar și cu 35° Celsius) în primele două veri, și cu jumătate din această valoare în întreg deceniul următor.*

* Injectarea carbonului în atmosferă poate provoca două tipuri de modificare a climei: încălzire, din cauza dioxidului de carbon, sau răcire, din cauza fumului și funinginei. Nu doar primul tip e contestat fără vreo dovadă științifică: mi se spune uneori că s-a demonstrat că iarna nucleară e imposibilă. Răspund de fiecare dată cerând să mi se indice un articol dintr-o revistă științifică serioasă unde se fac asemenea afirmații tranșante, iar deocamdată se pare că nu există așa ceva. Deși persistă mari incertitudini și e nevoie de noi cercetări, în special în legătură cu cantitatea de fum produsă și altitudinea la care acesta se ridică, opinia mea științifică este că nu avem nici un temei pentru a ignora riscul iernii nucleare. (N. a.)

Ce înseamnă de fapt asta? Nu-i nevoie să ai prea multe cunoștințe agricole ca să-ți dai seama că temperaturi aproape de punctul de îngheț în timpul verii ani de-a rândul ar distruge majoritatea producției de hrană. Este greu de prevăzut cu exactitate ce s-ar întâmpla după ce mii dintre cele mai mari orașe ar fi transformate în moloz și infrastructura globală s-ar prăbuși, iar puținii oameni care ar supraviețui foametei, hipotermiei sau bolilor ar trebui să înfrunte bande înarmate căutând cu disperare hrană.

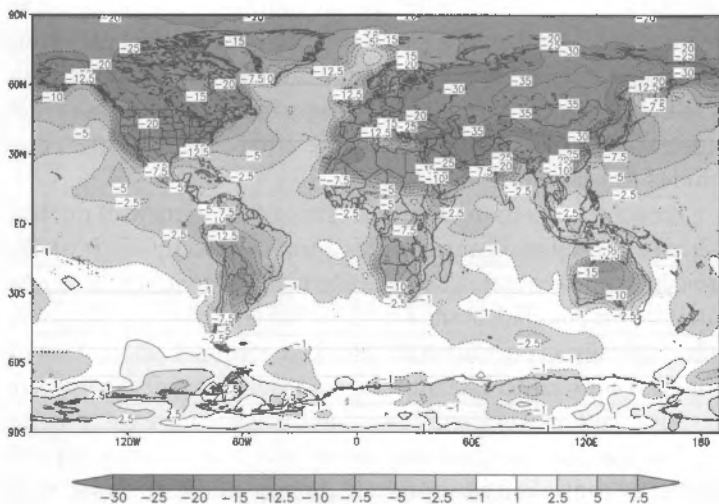


Figura 5.3 Răcirea medie (în °C) în cursul primelor două veri după un război nuclear total între Statele Unite și Uniunea Sovietică. Reprodus după Alan Robock *et al.*¹¹

Am intrat în detalii privind războiul nuclear global pentru a sublinia că deși nici un lider mondial rezonabil nu și-ar dori așa ceva, războiul s-ar putea declanșa în mod accidental. Nu putem avea așadar garanția că semenii noștri nu vor comite niciodată un omnicid: pentru a împiedica asta nu-i suficient ca nimeni să n-o vrea.

Arma apocalipsei

Am putea oare scăpa dintr-un omnicid? Chiar dacă un război nuclear poate ucide 90% din oameni, majoritatea savanților bănuiesc că nu ne-ar ucide pe toți, deci nu s-ar ajunge la extincție. Pe de altă parte, scenariile legate de radiațiile nucleare, EMP nuclear și iarna nucleară demonstrează că cele mai mari riscuri pot fi unele la care nici măcar nu ne-am gândit. Este extrem de greu să prevezi toate urmările și felul în care iarna nucleară, distrugerea infrastructurii, nivelurile ridicate ale mutațiilor și hoardele înarmate ar putea interacționa cu alte aspecte, cum sunt noi pandemii, prăbușirea ecosistemului și efecte pe care încă nu ni le imaginăm. Cred că deși probabilitatea ca mâine să izbucnească un război nuclear care să ducă la extincția oamenilor e mică, nu putem conchide ferm că e nulă.

Probabilitatea omnicidului crește dacă armamentul nuclear actual e perfecționat pentru se crea o armă a apocalipsei (*doomsday device*). Definită ca atare în 1960 de Herman Kahn, strateg la compania RAND, și popularizată de filmul lui Stanley Kubrik *Dr. Strangelove*, arma apocalipsei duce ideea de distrugere reciprocă garantată la ultimele ei consecințe. Este intimidarea absolută: un dispozitiv care răspunde automat oricărui atac inamic ucigând întreaga omenire.

Un candidat pentru arma apocalipsei este un uriaș depozit subteran conținând așa-numitele „bombe nucleare sărate”*, de preferință bombe cu hidrogen înconjurate de cantități mari de cobalt. Fizicianul Leo Szilard susținea încă din 1950 că în felul acesta ar putea fi uciși toți oamenii de pe Pământ: exploziile bombelor cu hidrogen ar face să devină radioactiv cobaltul și l-ar ejecta în stratosferă, iar timpul de înjumătățire de cinci ani al cobaltului e suficient de lung pentru ca acesta să se depună pe întreg Pământul (mai ales dacă arme apocaliptice gemene ar fi amplasate în emisfera opusă), dar suficient de scurt pentru ca

* *Salted nukes* – bombe a căror explozie generează cantități mari de substanțe radioactive care se depun pe sol, făcând nelocuibile suprafețe întinse. Se numesc „sărate” prin analogie cu sărăturile, terenuri care conțin un exces de săruri minerale și sunt greu cultivabile. (N. I.)

intensitatea radiației să fie letală. Există informații că au început deja să fie construite bombe cu cobalt. Riscul omnicidului poate spori dacă se adaugă bombe optimizate pentru declanșarea iernii nucleare prin creșterea la maximum a cantității de aerosoli cu timpi lungi de viață în stratosferă. Avantajul unei asemenea arme a apocalipsei este că e mult mai ieftină decât alte mijloace nucleare de intimidare: din moment ce nu-i nevoie ca bombele să fie lansate, nu sunt necesare sisteme costisitoare de rachete, iar bombele însele sunt mai ieftine, fiindcă nu trebuie să fie ușoare și compacte pentru a fi transportate de rachete.

Altă posibilitate este crearea în viitor a unei arme biologice a apocalipsei: o bacterie sau un virus care să ucidă toți oamenii. Dacă se transmite ușor și perioada de incubație e suficient de lungă, practic toată lumea ar fi infectată înainte să i se descopere existența și să se ia contramăsuri. Există un argument de ordin militar pentru construirea unei asemenea arme biologice, chiar dacă nu poate ucide toți oamenii: cea mai eficientă armă a apocalipsei combină toate tipurile de armament (nuclear, biologic etc.) pentru a spori la maximum șansele de descurajare a inamicului.

Arme IA

A treia cale tehnologică spre omnicid poate implica arme IA relativ simple. Să presupunem că o superputere construiește miliarde de drone de dimensiunea unei albine, ca acelea despre care am vorbit în capitolul 3, destinate să ucidă pe toată lumea, cu excepția cetățenilor proprii și ai aliaților, identificați de la distanță printr-un semnal de radiofrecvență, așa cum se întâmplă azi cu etichetele de pe produsele din supermarketuri. Aceste etichete ar fi distribuite tuturor cetățenilor pentru a fi purtate pe brățări sau ar fi implantate subcutanat, ca în discuția despre totalitarism. Asta ar determina probabil o superputere inamică să construiască ceva analog. Când războiul s-ar declanșa accidental, toată omenirea ar fi ucisă, chiar și triburile izolate, fiindcă nimeni nu poartă ambele tipuri de etichete de identificare. În combinație cu o armă nucleară și biologică a apocalipsei, riscul omnicidului ar fi și mai mare.

La începutul acestui capitol te-am invitat să te gândești unde ai vrea să conducă actuala cursă către IAG. Acum, că am explorat împreună o gamă largă de scenarii, care dintre ele te atrage, și pe care crezi că ar trebui să încercăm din răputeri să-l evităm? Ai un favorit clar? Dă-ne de știre mie și celorlalți cititori la <http://AgeOfAi.org>, și alătură-te dezbaterii noastre!

Evident, scenariile explorate nu alcătuiesc o listă completă, iar multe sunt tratate fără a intra în detalii, însă m-am străduit să cuprind cât mai mult, acoperind întregul spectru, de la tehnologie înaltă la tehnologie modestă și absența tehnologiei, și prezentând principalele speranțe și temeri exprimate în literatura de specialitate.

Scriind cartea de față, m-a interesat să aflu ce cred prietenii și colegii mei despre aceste scenarii, și am descoperit că nu există vreun consens. Singurul lucru asupra căruia cad de acord este că alegerile sunt mai subtile decât par la prima vedere. Cei care preferă un scenariu anume îi găsesc de regulă și neajunsuri. Asta înseamnă, cred eu, că trebuie să continuăm și să aprofundăm dezbaterea privind obiectivele noastre viitoare, pentru a ști care este direcția de urmat. Potențialul viitor al vieții în cosmos e măreț, și n-are rost să-l risipim navigând fără busolă, fără indicii privind ținta noastră.

Dar cât de măreț este acest potențial viitor? Oricât ar progresa tehnologia noastră, capacitatea Vieții 3.0 de a se perfecționa și a se răspândi în cosmos va fi limitată de legile fizicii – care sunt aceste limite absolute, în cursul următoarelor miliarde de ani? Forfotește acum universul nostru de viață extraterestră, sau suntem singuri? Ce se întâmplă când se întâlnesc diferite civilizații cosmice aflate în expansiune? Vom aborda aceste întrebări fascinante în următorul capitol.

REZUMAT:

- Actuala cursă către IAG poate conduce către un spectru larg de scenarii pentru următoarele milenii.
- Superintelența poate coexista pașnic cu oamenii fie pentru că e obligată (scenariul zeu înrobît), fie pentru că este o „IA prietenoasă” care își dorește asta (scenariile utopie libertariană, zeu protector, dictator binevoitor și paznic la zoo).
- Apariția superintelenței poate fi împiedicată de o IA (scenariul gardian) sau de oameni (scenariul 1984), de uitarea voluntară a tehnologiei (scenariul revenire) sau de absența motivației de a o construi (scenariul utopie egalitară).
- Omenirea poate dispărea, fiind înlocuită de inteligențe artificiale (scenariile învingători și descendenți) sau de nimic (scenariul autodistrugere).
- Nu există un consens în privința scenariului de dorit, toate implică elemente criticabile. Este deci cu atât mai important să continuăm și să aprofundăm dezbaterile despre obiectivele noastre viitoare, pentru a nu ne îndrepta sau rătăci într-o direcție nefericită.

CAPITOLUL 6

Zestrea noastră cosmică: următorul miliard de ani și mai târziu

Speculația noastră conduce către o supercivilitație, sinteza întregii vieți din sistemul solar, perfecționându-se și extinzându-se constant, răspândindu-se dincolo de Soare, transformând non-viața în minte — Hans Moravec, *Mind Children*

Pentru mine, cea mai fascinantă descoperire științifică din toate timpurile este aceea că am subestimat enorm potențialul viitor al vieții. Visurile și aspirațiile noastre nu trebuie să se limiteze la durate de viață întinse pe un secol, apăsate de boli, sărăcie și durată. Ajutată de tehnologie, viața are posibilitatea să înflorească vreme de miliarde de ani, nu doar în sistemul nostru solar, dar și în cosmos, cu o măreție pe care strămoșii noștri nu și-o puteau închipui. Nici măcar cerul nu e limita.

E o veste extraordinară pentru o specie care, de-a lungul epocilor, și-a dorit mereu să depășească limite. Jocurile Olimpice glorifică depășirea limitelor puterii, vitezei, agilității și rezistenței. Știința glorifică depășirea limitelor cunoașterii și înțelegerii. Literatura și arta glorifică depășirea limitelor în crearea frumosului sau în transmiterea unor trăiri care îți îmbogățesc viața. Oameni, organizații și națiuni glorifică sporirea resurselor, teritoriului și longevității. Dată fiind obsesia noastră pentru limite, nu-i de mirare că cea mai bine vândută dintre cărțile cu drepturi de autor este *The Guinness Book of World Records*.

Așadar, dacă limitele vieții pe care ni le închipuiam în trecut pot fi spulberate de tehnologie, care sunt limitele *ultime*? Cât din cosmosul nostru poate prinde viață? Cât de departe poate ajunge

viața și cât de mult poate dăinui? Cât de multă materie poate folosi viața, și cât de multă energie, informație și putere de calcul poate ea extrage. Aceste limite ultime nu sunt impuse de înțelegerea noastră, ci de legile fizicii. Paradoxal, asta face ca, în anumite privințe, viitorul pe termen lung să fie mai ușor de analizat decât viitorul pe termen scurt.

Dacă cei 13,8 miliarde de ani de istorie ar fi comprimați într-o săptămână, atunci drama întinzându-se pe 10 000 de ani din ultimele două capitole s-ar sfârși în mai puțin de o jumătate de secundă. Ca atare, deși nu putem prezice dacă și cum va avea loc o explozie a inteligenței, și care vor fi urmările ei imediate, toată această agitație este doar un scurt moment din istoria cosmică, ale cărui detalii nu influențează limitele ultime ale vieții. Dacă viața post-explozie e în aceeași măsură obsedată de depășirea limitelor ca oamenii de azi, atunci ea va elabora tehnologia prin care să atingă efectiv aceste limite ultime – pentru că o poate face. În capitolul de față vom vedea care sunt limitele, făcându-ne astfel o idee despre cum ar putea arăta viitorul pe termen lung al vieții. Aceste limite se bazează pe cunoștințele noastre actuale de fizică, deci trebuie privite ca o eventuală subestimare a posibilităților: descoperiri științifice ulterioare ar putea sugera căi de depășire a lor.

Știm însă cu certitudine că viața viitoare va fi atât de ambițioasă? Nu, nu știm: poate că va deveni la fel de autosatisfăcută ca un consumator de heroină sau ca un snob care nu face decât să privească la nesfârșit reluări ale episodelor din *Keeping Up with the Kardashians*.^{*} Avem totuși motive să bănuim că ambiția e o trăsătură universală a vieții avansate. Aproape indiferent ce încercă să sporească la maximum – inteligență, longevitate, cunoaștere sau experiențe interesante –, va avea nevoie de resurse. Va fi deci stimulată să împingă tehnologia până la limitele ultime, pentru a obține cât mai mult din resursele de care dispune. Apoi, singura cale de a merge mai departe este să dobândească noi resurse, extinzându-se în regiuni tot mai mari din cosmos.

* Reality-show american care prezintă viața familiei Kardashian. (N. t.)

Viața ar putea însă apărea în mod independent în mai multe locuri din univers. În acest caz, civilizațiile neambicioase ar deveni pur și simplu irelevante cosmic, părți din ce în ce mai mari din zestrea cosmică fiind în cele din urmă luate în stăpânire de forme de viață mai ambițioase. Selecția naturală se manifestă deci la scară cosmică, iar după o vreme aproape toate formele de viață existente vor fi ambițioase. Pe scurt, dacă vrem să aflăm în ce măsură cosmosul nostru poate căpăta viață în cele din urmă, trebuie să cercetăm limitele pe care legile fizicii le impun asupra ambiției. Să explorăm mai întâi limitele a ceea ce putem face cu resursele (materie, energie etc.) de care dispunem în sistemul nostru solar, pentru a vedea apoi cum putem obține mai multe resurse prin explorare și colonizare cosmică.

FOLOSIREA LA MAXIMUM A RESURSELOR DE CARE DISPUI

Dacă supermarketurile și bursele de mărfuri de azi vând zeci de mii de produse pe care le-am putea numi „resurse”, viața din viitor care a atins limitele tehnologice are în principal nevoie de o resursă fundamentală: așa-numita *materie barionică*, adică tot ce e alcătuit din atomi sau constituenții lor (cuarci și electroni). Indiferent sub ce formă se găsește această materie, tehnologia avansată o poate rearanja în orice substanțe sau obiecte se dorește – de pildă, în centrale electrice, calculatoare sau forme superioare de viață. Să începem deci prin a examina limitele impuse asupra energiei care alimentează viața avansată și asupra procesării informației care îi permite să gândească.

Construirea sferelor Dyson

Dacă vine vorba despre viitorul vieții, unul dintre cei mai optimiști vizionari este Freeman Dyson. Am onoarea și plăcerea să-l cunosc de două decenii, dar la prima noastră întâlnire am fost foarte emoționat. Eram un tânăr cercetător postdoctoral și mâncam împreună cu

prietenii mei la cantina Institutului pentru Studii Avansate de la Princeton, când brusc a apărut acest celebru fizician, care pe vremuri obișnuia să stea de vorbă cu Einstein și Gödel, s-a prezentat și ne-a întrebat dacă poate să ni se alăture! M-a făcut însă imediat să mă simt în largul meu, explicându-ne că preferă să ia prânzul cu tinerii decât cu bătrânii profesori pedanți. Deși în momentul în care scriu aceste cuvinte a ajuns la 93 de ani, Freeman are un spirit mai tânăr decât majoritatea celor pe care-i cunosc, iar lucirea poznașă a ochilor săi arată cât de puțin îi pasă de formalități, ierarhii academice sau înțelepciunea convențională. Cu cât ideea e mai îndrăznească, cu atât se pasionează mai tare.

Când am discutat despre folosirea energiei, a ridiculizat lipsa de ambiție omenească, arătând că am putea satisface toate nevoile noastre energetice actuale prin captarea luminii solare care ajunge pe o suprafață mai mică de 0,5% din Sahara. Dar de ce să ne oprim aici? De ce să ne oprim chiar și la captarea întregii lumini solare care ajunge pe Pământ, lăsând să se irosească în spațiul gol cea mai mare parte a ei? De ce să nu punem pur și simplu toată producția de energie a Soarelui în folosul vieții?

Inspirat de romanul SF *Star Maker* al lui Olaf Stapledon din 1937, cu inele de lumi artificiale orbitând în jurul stelei lor, Freeman Dyson a publicat în 1960 descrierea a ceea ce avea să fie cunoscut drept *sfera Dyson*.¹ Ideea lui era de a rearanja planeta Jupiter într-o biosferă sub forma unui strat sferic înconjurând Soarele, în care descendenții noștri s-ar dezvolta având la dispoziție de 100 de miliarde de ori mai multă biomasă și de un bilion de ori mai multă energie decât folosește omenirea în prezent.² Dyson susținea că acesta e următorul pas firesc: „Este de așteptat ca, la câteva mii de ani după ce a intrat în epoca industrială, orice specie inteligentă să ocupe o biosferă artificială care înconjoară complet steaua ei.” Dacă locuiești în interiorul unei sfere Dyson, nu există nopți, vezi venind de pretutindeni lumina Soarelui reflectată de restul biosferei, așa cum vezi azi lumina Soarelui reflectată de Lună. Dacă vrei să vezi stelele, trebuie să „urci” și să privești cosmosul din afara sferei Dyson.

O soluție presupunând tehnologie de nivel mai modest pentru a construi o sferă Dyson parțială este plasarea unui inel de habitate pe o orbită circulară în jurul Soarelui. Pentru a înconjura complet Soarele, adaugi inele sub axe diferite și la distanțe ușor diferite, ca să eviți ciocnirile. Dificultatea de a conecta între ele aceste inele aflate în mișcare rapidă, ceea ce complică transportul și comunicarea, poate fi depășită dacă se construiește o sferă Dyson staționară monolitică, la care atracția gravitațională a Soarelui e echilibrată de presiunea radiației solare – idee propusă de Robert L. Forward și Colin McInnes. Sfera poate fi construită adăugând treptat „sateliți”: sateliți staționari care contrabalansează gravitația Soarelui prin presiunea radiației, nu prin forțe centrifuge. Ambele forțe scad cu pătratul distanței până la Soare, ceea ce înseamnă că dacă pot fi echilibrate la o distanță față de Soare, ele se vor echilibra la orice distanță, sateliții putând fi plasați oriunde în sistemul solar. Sateliții trebuie să fie niște foi extrem de ușoare, cântărind doar 0,77 grame de metru pătrat, de vreo sută de ori mai puțin decât hârtia, dar asta nu constituie un impediment. De exemplu, o foaie de grafen (un singur strat de atomi de carbon aranjați într-o structură hexagonală de tip fagure) cântărește de o mie de ori mai puțin decât acea greutate-limită. Dacă sfera Dyson e construită așa încât să nu absoarbă, ci să reflecte cea mai mare parte a luminii solare, atunci crește intensitatea reflectării luminii, sporind presiunea radiației și făcând astfel posibil ca sfera să devină și mai grea. Multe stele au o luminozitate de o sută, ba chiar de o mie de ori mai mare decât Soarele nostru, deci pot suporta sfere Dyson cu atât mai grele.

Dacă vrem să avem în sistemul nostru solar o sferă Dyson rigidă mult mai grea, atunci pentru a se opune gravitației Soarelui e nevoie de materiale ultrarezistente care pot suporta presiuni de zeci de mii de ori mai mari decât acelea din structura de rezistență a celor mai înalți zgârie-nori, care nu se topesc și nu se deformează. Pentru a dăinui multă vreme, o sferă Dyson trebuie să fie dinamică și inteligentă, reglându-și constant poziția și forma ca răspuns la perturbații, și deschizând găuri largi pentru a lăsa asteroizii și cometele să treacă prin ea fără incidente. Se poate la fel de bine

ca un sistem de detecție și deviere să fie folosit pentru a face față acestor intruși, eventual dezmembrându-i și utilizându-le materia în mod convenabil.

Pentru oamenii de azi, viața pe o sferă Dyson sau în interiorul ei ar fi în cel mai bun caz derutantă, iar în cel mai rău caz imposibilă, dar asta nu va împiedica viitoarele forme de viață, biologice sau nebiologice, să se dezvolte acolo. În varianta cu sateliți, practic n-ar exista gravitație, iar dacă te-ai plimba pe un satelit staționar ai merge doar pe suprafața orientată spre exterior (în direcția opusă Soarelui) fără să cazi, sub acțiunea unei gravitații de vreo zece mii de ori mai slabă decât cea cu care ești obișnuit. N-ar exista câmp magnetic (cu excepția cazului în care tu îl creezi) și ai fi protejat de particulele periculoase emise de Soare. Avantajul ar fi că o sferă Dyson de dimensiunea orbitei actuale a Pământului ne-ar pune la dispoziție o suprafață de vreo 500 de milioane de ori mai mare decât cea pe care trăim azi.

Dacă ne dorim habitate cu condiții asemănătoare celor de pe Pământ, vestea bună este că ele sunt mult mai ușor de construit decât o sferă Dyson. De exemplu, figurile 6.1 și 6.2 prezintă planul unui habitat cilindric conceput de fizicianul american Gerard K. O'Neil, cu gravitație artificială, protecție contra radiației cosmice, ciclul zi-noapte de 24 de ore și atmosferă și ecosisteme de tipul celor de pe Pământ. Asemenea habitate ar putea orbita liber în interiorul unei sfere Dyson, sau variante modificate ar putea fi atașate în exteriorul ei.

Construirea unor centrale electrice mai eficiente

Deși sferile Dyson sunt eficiente energetic în raport cu standardele de azi, ele împing doar cu puțin limitele impuse de legile fizicii. Einstein ne-a arătat că dacă vrem să convertim masa în energie cu o eficiență de 100%,* atunci o cantitate de masă m ne-ar

* Dacă lucrezi în sectorul energetic ești pesemne obișnuit să definești eficiența drept fracțiunea din energia eliberată care se află într-o formă utilă. (N. a.)

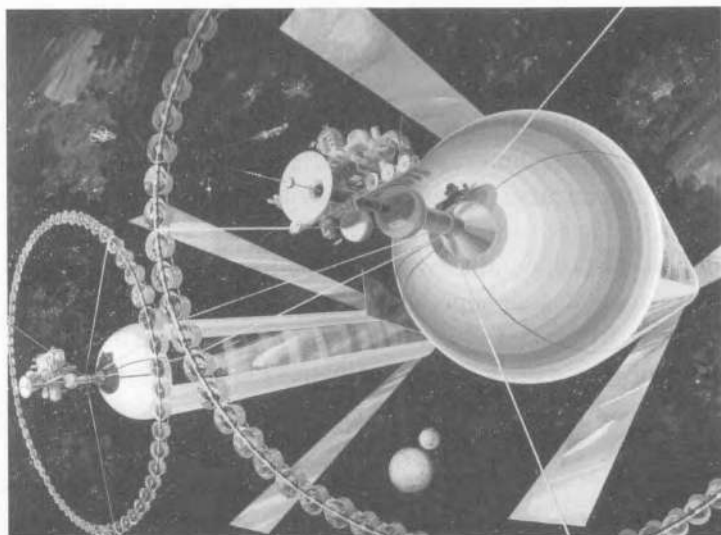


Figura 6.1 O pereche de cilindri O'Neil în contrarotație ar oferi habitate confortabile de tip terestru dacă ar orbita în jurul Soarelui astfel încât să fie orientate mereu drept spre el. Forța centrifugă a rotației lor produce o gravitație artificială, iar trei oglinzi pliabile proiectează în interior lumina solară într-un ciclu zi-noapte de 24 de ore. Habitatele mici dispuse într-un inel sunt destinate agriculturii. Imagine obținută prin amabilitatea lui Rick Guidice/NASA.

da o cantitate de energie E conform celebrei formule $E = mc^2$, unde c este viteza luminii. Asta înseamnă că, c fiind foarte mare, o cantitate mică de masă poate produce o cantitate uriașă de energie. Dacă am avea rezerve importante de antimaterie (dar n-avem), atunci ar fi ușor de construit o centrală electrică 100% eficientă: turnând o linguriță de anti-apă în apă obișnuită s-ar elibera energie echivalentă cu cea a 200 000 de tone de TNT, corespunzând unei bombe cu hidrogen tipice – suficientă pentru a satisface nevoile de energie ale întregii lumi pentru circa șapte minute.

În prezent însă, cele mai răspândite modalități de generare a energiei sunt jalnic de ineficiente, după cum se vede în tabelul 6.1 și în figura 6.3. Digerarea unei bomboane are o eficiență de doar 0,00000001%, în sensul că eliberează doar o zecime de biliomime din energia mc^2 pe care o conține. Dacă stomacul tău ar

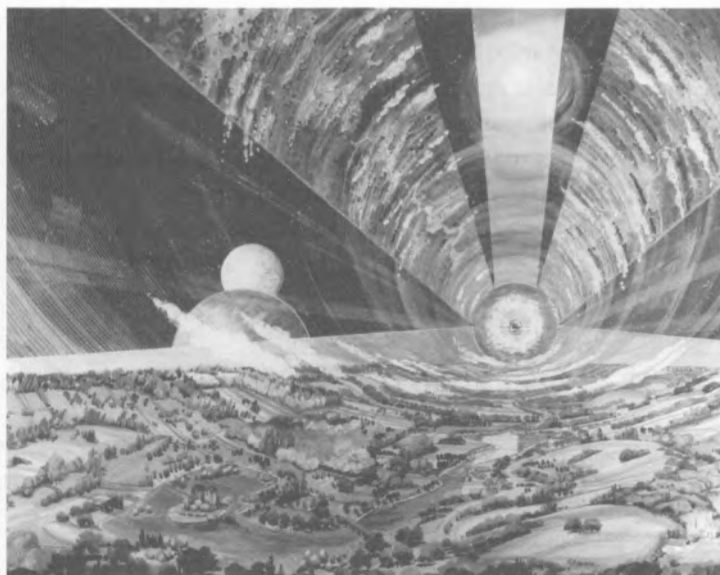


Figura 6.2 Interiorul unuiu dintre cilindrii O'Neil din figura precedentă. Dacă diametrul lui este de 6,4 km și efectuează o rotație la fiecare 2 minute, oamenii vor simți aceeași gravitație aparentă ca pe Pământ. Soarele se află în spate, dar pare să fie deasupra grație unei oglinzi din exteriorul cilindrului care se pliază noaptea. Ferestre ermetice împiedică atmosfera să iasă din cilindru. Imagine obținută prin amabilitatea lui Rick Guidice/NASA.

avea o eficiență chiar și de 0,001%, ai avea nevoie să mănânci o singură dată pentru tot restul vieții. În comparație cu mâncatul, arderea cărbunelui și benzinei este de doar 3, respectiv 5 ori mai eficientă. Reactoarele nucleare de azi se descurcă mult mai bine despicând atomii de uraniu prin fisiune, dar tot nu reușesc să extragă mai mult de 0,08% din energia lor. Reactorul nuclear din miezul Soarelui este cu un ordin de mărime mai eficient, extrăgând 0,7% din fuziunea hidrogenului pentru a produce heliu. Chiar dacă am înconjura Soarele cu o sferă perfectă, am converti abia 0,08% din masa Soarelui în energie pe care s-o putem folosi, fiindcă atunci când Soarele își va consuma circa o zecime din hidrogenul său, își va încheia viața de stea normală, devenind o gigantă roșie și începând să moară. Lucrurile nu stau mai bine

nici pentru alte stele: fracțiunea de hidrogen consumată în decursul vieții lor principale variază între 4%, la stelele foarte mici, și 12%, la stelele cele mai mari. Dacă am construi un reactor de fuziune care să folosească 100% din hidrogenul pe care îl avem la dispoziție, am rămâne tot la eficiența de 0,7% a procesului de fuziune. Cum ne-am putea descurca mai bine?

Metodă	Eficiență
Digerarea unei bomboane	0,00000001%
Arderea cărbunelui	0,00000003%
Arderea benzinei	0,00000005%
Fisiunea uraniului-235	0,08%
Folosirea unei sfere Dyson până când Soarele moare	0,08%
Fuziunea hidrogenului pentru a produce heliu	0,7
Motor cu gaură neagră în rotație	29%
Sferă Dyson în jurul unui quasar	42%
Sfalerizor	50%?
Evaporarea găurilor negre	90%

Tabelul 6.1 Eficiența convertirii masei în energie utilizabilă, în raport cu limita teoretică $E = mc^2$. După cum se arată în text, procesul în care hrănești găurile negre și aștepti ca ele să se evapore, cu o eficiență energetică de 90%, este din păcate prea lent pentru a fi util, iar accelerarea procesului îi scade mult eficiența.

Evaporarea găurilor negre

În cartea sa *O scurtă istorie a timpului*, Stephen Hawking propune ideea de centrală electrică pe bază de găuri negre.* Poate părea paradoxal, odată ce multă vreme s-a crezut că găurile negre sunt capcane din care nimic nu poate vreodată evada, nici măcar lumina. Hawking însă a calculat că efectele gravitației cuantice fac ca o gaură neagră să se comporte ca un obiect fierbinte – cu cât

* Dacă în universul apropiat nu se află nici o gaură neagră naturală potrivită, poți crea una înghesuind o cantitate mare de materie într-un spațiu suficient de mic. (N. a.)

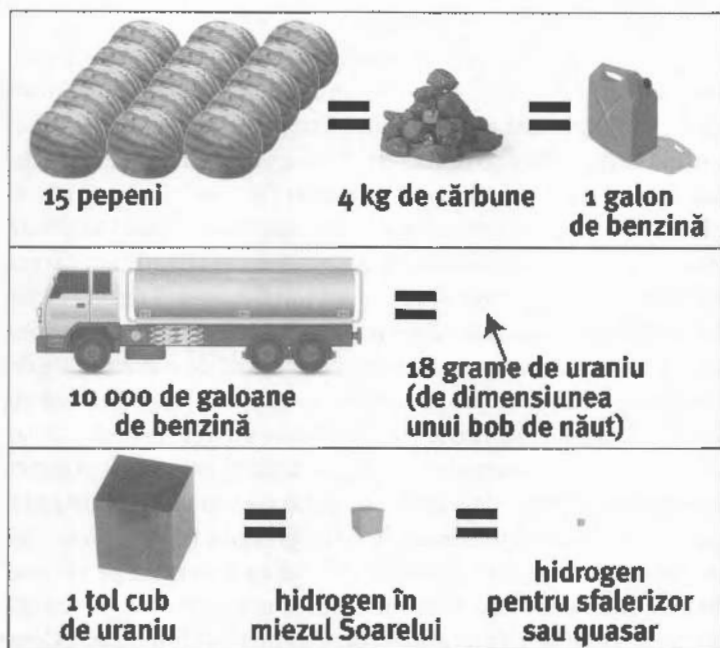


Figura 6.3 Tehnologia avansată poate extrage mult mai multă energie din materie decât obținem mâncând sau arzând-o, iar până și fuziunea extrage de 140 de ori mai puțină energie decât am extrage dacă ne-am situa la limitele impuse de fizică. Centrale electrice exploatând sfaleroni, quasari sau găuri negre care se evaporă ar fi mult mai eficiente.

mai mic, cu atât mai fierbinte – care emite o radiație termică numită acum *radiație Hawking*. Asta înseamnă că, treptat, gaura neagră pierde energie și se evaporă. Cu alte cuvinte, indiferent ce materie arunci în gaura neagră, ea se va întoarce sub formă de radiație termică, așa încât, atunci când gaura neagră s-a evaporat complet, ai convertit materia în radiație cu o eficiență de aproape 100%.*

* Este o supraestimare, fiindcă radiația Hawking conține și particule de la care e greu de extras lucru mecanic util. Găurile negre mari au o eficiență de doar 90%, pentru că aproximativ 10% din energie e radiată sub formă de gravitoni: particule extrem de „timide”, care sunt aproape imposibil de detectat, ca să nu mai vorbim de extragerea lucrului mecanic util. Pe măsură ce gaura neagră continuă să se evapore și să se mișcoreze, eficiența scade și

Dacă nu cumva gaura neagră are dimensiuni mai mici decât un atom, este problematică folosirea evaporării ei ca sursă de energie, pentru că procesul e neverosimil de lent – durează mai mult decât vârsta actuală a universului –, iar gaura neagră radiază mai puțină energie decât o lumânare. Puterea produsă scade cu pătratul dimensiunii găurii negre, așa încât fizicienii Louis Crane și Shawn Westmoreland au propus folosirea unei găuri negre de vreo mie de ori mai mică decât un proton, cântărind cam cât cel mai greu vapor.³ Ideea lor era ca motorul cu gaură neagră să fie folosit la propulsia unei nave spațiale (subiect asupra căruia vom reveni), așa încât erau preocupați mai curând de portabilitate decât de eficiență, și au propus ca gaura neagră să fie alimentată cu lumină laser, fără să existe vreo conversie energie-materie. Chiar dacă ai alimenta-o cu materie în loc de radiație, este dificil să garantezi obținerea unei eficiențe înalte: pentru a intra într-o gaură neagră de o mie de ori mai mică decât ei, protonii trebuie proiectați asupra găurii negre cu o instalație la fel de puternică pe cât este Marele Accelerator de Hadroni*, crescând energia lor mc^2 cu cel puțin de o mie de ori mai multă energie cinetică (de mișcare). Cum cel puțin 10% din acea energie cinetică s-ar pierde prin gravitoni atunci când gaura neagră se evaporă, am introduce mai multă energie în gaura neagră decât am extrage, obținând o eficiență negativă. Altă piedică în calea construirii unei centrale cu găuri negre ține de faptul că nu avem încă o teorie riguroasă a gravitației cuantice pe care să ne bazăm calculele – dar această incertitudine lasă deschisă posibilitatea să existe efecte utile ale gravitației cuantice pe care deocamdată nu le-am descoperit.

Găuri negre rotitoare

Există din fericire alte căi de a folosi găurile negre în centrale, fără a implica gravitația cuantică sau alte aspecte puțin înțelese ale

mai mult, fiindcă radiația Hawking începe să includă neutrini și particule masive. (N. a.)

* Large Hadron Collider (LHC) – accelerator de particule construit la Centrul European pentru Cercetări Nucleare (CERN), în apropiere de Geneva. (N. t.)

fizicii. De pildă, multe găuri negre se învârt foarte repede în jurul axei proprii, orizontul evenimentelor rotindu-se cu o viteză apropiată de cea a luminii, iar această energie de rotație poate fi extrasă. Orizontul evenimentelor unei găuri negre este regiunea din care nici măcar lumina nu poate evada, pentru că atracția gravitațională e prea puternică. Figura 6.4 arată că, pentru o gaură neagră în rotație, în afara orizontului evenimentelor există o regiune numită *ergosferă*, în care gaura neagră antrenează în mișcare spațiul atât de repede, încât nici o particulă nu poate rămâne în repaus, fiind și ea antrenată în mișcare. Dacă arunci un obiect în

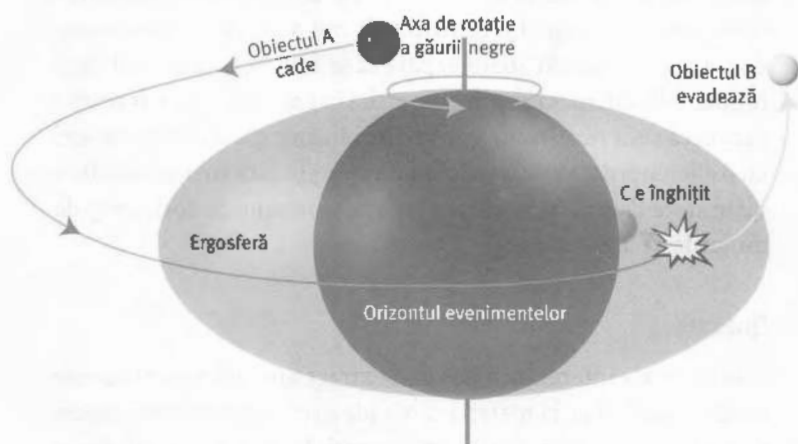


Figura 6.4 O parte din energia de rotație a unei găuri negre care se învârt în jurul axei proprii poate fi extrasă aruncând o particulă A în apropierea găurii negre și făcând-o să se rupă într-o parte C care e înghițită și o parte B care evadează – cu energie mai mare decât a avut inițial A.

ergosferă, acesta va fi accelerat într-o mișcare rapidă în jurul găurii negre. Din păcate, va fi curând înghițit de ea, dispărând pentru totdeauna prin orizontul evenimentelor, așadar nu poți extrage astfel energie. Totuși, Roger Penrose a descoperit că dacă lansezi obiectul sub un anumit unghi și faci așa încât să se rupă în două bucăți, după cum se vede în figura 6.4, se poate ca doar o bucată să fie înghițită, iar cealaltă să se îndepărteze de gaura

neagră cu o energie mai mare decât cea cu care ai lansat obiectul inițial. Altfel spus, ai convertit o parte din energia de rotație a găurii negre în energie pe care o poți folosi. Repetând de multe ori procesul, poți extrage *toată* energia de rotație a găurii negre, așa încât aceasta încetează să se rotească, iar ergosfera ei dispare. Dacă gaura neagră inițială se rotește pe cât de repede permit legile naturii, orizontul evenimentelor deplasându-se practic cu viteza luminii, prin această strategie ai putea converti în energie 29% din masa ei. Nu se știe încă precis cât de repede se rotesc găurile negre observate, dar multe dintre cele atent studiate par să se rotească foarte repede: între 30% și 100% din maximumul permis. Gaura neagră uriașă din centrul galaxiei noastre (de 4 milioane de ori mai grea decât Soarele) pare să se rotească, deci chiar dacă numai 10% din masa ei ar fi convertită în energie utilă, s-ar obține cam ceea ce ar rezulta din convertirea în energie a 400 000 de sori cu o eficiență de 100%, sau tot atâta energie câtă am extrage de la niște sfere Dyson înconjurând 500 de milioane de sori, timp de miliarde de ani.

Quasari

Altă strategie interesantă este de a extrage energie nu de la gaura neagră însăși, ci de la materia care cade în ea. Natura a găsit deja o modalitate de a face asta – prin quasari. Pe măsură ce un gaz se rotește în apropierea unei găuri negre, formând un disc ale cărui părți centrale sunt treptat înghițite, gazul devine extrem de fierbinte și emite mari cantități de radiație. Când gazul cade spre gaura neagră, e accelerat, transformând energia sa potențială gravitațională în energie cinetică. Turbulențe complexe fac ca mișcarea coordonată a masei de gaz să devină, la scară mică, tot mai dezordonată, până când atomii individuali încep să se ciocnească între ei la viteze mari – agitația aceasta înseamnă temperatură ridicată, iar ciocnirile violente convertesc energia de mișcare în radiație. Dacă am construi o sferă Dyson în jurul întregii găuri negre, la o distanță sigură, această energie a radiației ar putea fi captată și

folosită. Cu cât se rotește mai rapid gaura neagră, cu atât procesul devine mai eficient, la viteza de rotație maximă a găurii negre eficiența ajungând la 42%.* Pentru găurile negre de greutatea unei stele, majoritatea energiei e eliberată sub formă de raze X, iar pentru găurile negre supermasive, de genul celor din centrul galaxiilor, o mare parte a energiei se află în domeniile luminii infraroșii, vizibile și ultraviolete.

Când nu mai ai cu ce alimenta gaura neagră, poți începe să extragi energia ei de rotație, așa cum am văzut mai sus.** Și în acest caz natura a găsit deja o soluție, intensificând radiația unei mase de gaz printr-un proces magnetic numit mecanism Blandford-Znajek. Nu-i exclus ca tehnologia să ne ajute să obținem o eficiență în extragerea energiei de peste 42%, folosind în mod inteligent câmpuri magnetice și alte elemente.

Sfaleroni

Există și o modalitate de a converti materia în energie fără a implica găurile negre, printr-un proces implicând *sfaleronii*. Acesta poate distruge cuarzii, transformându-i în leptoni: electroni, miuoni și particule tau (verișorii mai grei ai electronilor), neutrini, precum și antiparticulele corespunzătoare tuturor acestora.⁴ După cum se arată în figura 6.5, modelul standard al fizicii particulelor prezice

* Fanii lui Douglas Adams ar trebui să observe că aceasta e o întrebare elegantă al cărei răspuns coincide cu cel la întrebarea privind viața, universul și tot ce există. Mai precis, eficiența este $1 - \sqrt{3} \approx 42\%$. (N. a.) [În seria de romane a lui Douglas Adams *Ghidul autostopistului galactic*, răspunsul ultim dat de un supercalculator este pur și simplu numărul 42, dar supercalculatorul nu mai poate spune apoi și care e întrebarea ultimă – n.t.]

** Dacă alimentezi gaura neagră plasând în jurul ei un nor de gaz care se rotește lent în același sens cu ea, atunci gazul se va învârti tot mai repede și va fi înghițit de ea, accelerând rotația găurii negre, la fel cum un patinator se învârtte mai repede când își apropie brațele de corp. În felul acesta rotația găurii negre poate fi menținută la viteză maximă, permițându-ți să extragi întâi 42% din energia gazului, iar apoi 29% din ce rămâne, eficiența totală fiind $42\% + (1 - 42\%) \times 29\% = 59\%$. (N. a.)

că nouă cuarci cu anumite valori ale aromei și spinului* se pot uni și transforma în trei leptoni, trecând printr-o stare intermediară numită sfaleron. Pentru că masa totală a ceea ce intră în reacție este mai mare decât masa totală a ceea ce rezultă, diferența de masă este convertită în energie conform formulei lui Einstein $E = mc^2$.

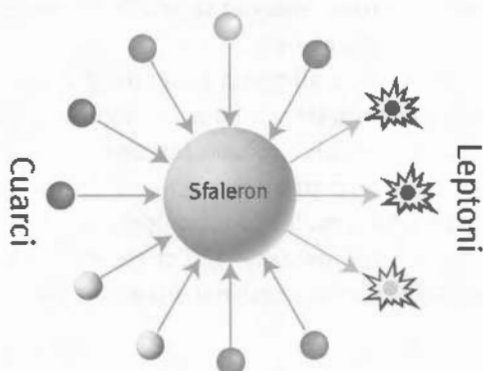


Figura 6.5 Conform modelului standard al fizicii particulelor, nouă cuarci cu anumite valori ale aromei și spinului se pot uni și transforma în trei leptoni, trecând printr-o stare intermediară numită sfaleron. Suma maselor cuarcilor (împreună cu energia gluonilor** care-i însoțesc) e mult mai mare decât suma maselor leptonilor, prin urmare în acest proces se va elibera energie, ilustrată grafic prin scântei.

Viața inteligentă din viitor ar putea deci construi ceea ce eu numesc un *sfalerizor*: un generator de energie acționând ca un motor diesel hipertrofiat. Motorul diesel tradițional comprimă un amestec de aer și motorină până când temperatura crește suficient pentru ca arderea să se declanșeze spontan, după care

* Spinul este o caracteristică a particulelor, asociată la începuturile mecanicii cuantice cu rotația particulei în jurul axei proprii (de la englezescul *to spin* – „a se învârti”). Aromă (*flavor*) este o caracteristică ce definește mai multe tipuri de cuarci și leptoni. (N. Ł)

** Gluonii (de la englezescul *glue* – „lipici”) sunt particulele care mijlocesc interacțiunile nucleare tari dintre cuarci, constituenți fundamentali ai materiei. (N. Ł)

amestecul fierbinte se destinde și efectuează lucru mecanic util – de pildă, împingând un piston. Dioxidul de carbon și alte gaze de combustie cântăresc cu aproximativ 0,00000005% mai puțin decât amestecul aflat inițial în cilindru, iar diferența de masă se transformă în energie termică ce acționează asupra pistonului. Un sfalerizor ar comprima materia obișnuită până la câteva trilioane (10^{15}) de grade, pentru a o lăsa apoi să se destindă și să se răcească, odată ce sfaleronii și-au îndeplinit misiunea.* Cunoaștem deja rezultatul acestui experiment, fiindcă universul a trecut prin așa ceva în urmă cu vreo 13,8 miliarde de ani, pe când era extrem de fierbinte: aproape 100% din materie s-a transformat în energie, rămânând mai puțin de o miliardime din particule pentru a alcătui materia obișnuită compusă din cuarci și electroni. Sfalerizorul seamănă deci cu un motor diesel, numai că e de peste un miliard de ori mai eficient! Alt avantaj este că nu contează ce combustibil introduci – funcționează cu orice e compus din cuarci, deci cu orice materie obișnuită.

Datorită acestor procese la temperatură înaltă, în universul nostru timpuriu a apărut de peste un bilion de ori mai multă radiație (fotoni și neutrini) decât materie (quarci și electroni, care mai târziu s-au agregat în atomi). În următorii 13,8 miliarde de ani s-a produs o netă separare, atomii s-au concentrat în galaxii, stele și planete, iar majoritatea fotonilor au rămas în spațiul intergalactic, formând radiația cosmică de fond de microunde, prin intermediul căreia am căpătat o imagine a universului timpuriu. Orice formă avansată de viață care trăiește într-o galaxie sau în altă concentrare de materie poate deci transforma cea mai mare parte din materia disponibilă înapoi în energie, readucând procentajul materiei la aceeași valoare minusculă de la începutul universului prin re-crearea pentru scurt timp a acelor condiții în interiorul unui sfalerizor.

* Materia trebuie să fie suficient de fierbinte pentru a reunifica forțele electromagnetice și nucleare slabă, ceea ce se întâmplă atunci când particulele se deplasează cu viteze corespunzând, într-un accelerador de particule, aplicării unei tensiuni de 200 de miliarde de volți. (N. a.)

Ca să ne dăm seama cât de eficient ar fi un sfalerizor, avem nevoie să cunoaștem unele detalii practice esențiale: de pildă, cât de mare trebuie să fie pentru a împiedica scurgerea unei fracțiuni importante din fotoni și neutrini în faza de compresie? Ceea ce putem totuși spune cu certitudine este că în viitor viața va beneficia de posibilități energetice incomparabil superioare celor permise de tehnologia actuală. N-am reușit nici măcar să construim un reactor de fuziune, dar tehnologia viitorului ar trebui să poată fi de zece, ba chiar de o sută de ori mai eficientă.

Calculatoare mai performante

Dacă eficiența energetică a digestiei este de 10 miliarde de ori mai mică decât limita fizică a convertirii masei în energie, cât de eficiente sunt calculatoarele de azi? După cum vom vedea, lucrurile stau încă și mai rău.

Spun adesea despre prietenul meu Seth Lloyd că e pesemne singurul coleg de la MIT care e la fel de nebun ca mine. După cercetări de pionierat în domeniul calculatoarelor cuantice, a scris o carte în care susținea că întregul univers e un calculator cuantic. După muncă, obișnuim să mergem la o bere, și n-am găsit încă un subiect despre care să nu aibă de spus ceva interesant. De pildă, așa cum am menționat în capitolul 2, are multe de spus despre limitele ultime ale calculului. Într-un celebru articol publicat în anul 2000, a demonstrat că viteza de calcul e limitată de energie: efectuarea unei operații logice elementare în timpul T necesită o energie medie $E = h/4T$, unde h este mărimea fundamentală din fizică numită constanta lui Planck. Asta înseamnă că 1 kilogram de calculator poate efectua cel mult 5×10^{50} operații pe secundă – cu 36 de ordine de mărime mai mult decât laptopul la care scriu aceste cuvinte! Vom atinge această limită în câteva secole, dacă puterea de calcul va continua să se dubleze la fiecare doi ani, după cum am văzut în capitolul 2. El a arătat de asemenea că 1 kilogram de calculator poate stoca cel mult 10^{31} biți, adică de circa un miliard de miliarde de ori mai mult decât laptopul meu.

Seth a fost primul care a înțeles că atingerea acestor limite ar fi dificilă chiar și pentru viața superinteligentă, fiindcă memoria aceluia 1 kilogram de „calculator” ultim ar semăna cu o explozie termonucleară sau cu un mic fragment din big bang. El crede însă că limitele practice nu sunt foarte departe de cele ultime. Prototipurile de calculatoare cuantice existente și-au miniaturizat deja memoria stocând un bit per atom, ceea ce la scară mare ar permite stocarea a circa 10^{25} biți/kg – de un bilion de ori mai mult decât laptopul meu. Apoi, folosind radiația electromagnetică pentru comunicarea între acești atomi, s-ar ajunge la aproximativ 5×10^{40} operații pe secundă – cu 31 de ordine de mărime mai mult decât unitatea centrală de procesare a laptopului meu.

Pe scurt, potențialul de calcul al vieții din viitor este într-adevăr uluitor: în termenii ordenelor de mărime, cele mai bune super-computere de azi se află mult mai departe de acel 1 kg de calculator ultim decât de semnalizatorul unei mașini, un dispozitiv care stochează doar un bit de informație, oscilând între conectat și deconectat o dată pe secundă.

Alte resurse

Din perspectiva fizicii, orice ar vrea să creeze viața viitoare – de la habitate și mașini la noi forme de viață – înseamnă pur și simplu aranjarea particulelor elementare într-un anumit mod. La fel cum o balenă albastră este mici crustacee rearanjate, iar micile crustacee sunt plancton rearanjat, întregul nostru sistem solar nu-i decât hidrogen rearanjat în cursul a 13,8 miliarde de ani de evoluție cosmică: gravitația a rearanjat hidrogenul în stele, care au rearanjat hidrogenul în atomi mai grei, după care gravitația a rearanjat acești atomi în planeta noastră, unde procesele chimice și biologice i-au rearanjat în viață.

Viața viitorului care și-a atins limitele tehnologice poate efectua asemenea rearanjări ale particulelor mult mai rapid și eficient, folosindu-și în primă instanță puterea de calcul pentru a găsi metoda optimă, iar apoi energia disponibilă pentru a alimenta procesul de rearanjare a materiei. Am văzut că materia poate fi

convertită atât în calculatoare, cât și în energie, așa încât, într-un anumit sens, este singura resursă fundamentală de care e nevoie.* Odată ce viața viitorului a atins limitele fizice pentru ceea ce poate face cu materia, îi rămâne o singură soluție de a face mai mult: să obțină mai multă materie. Iar pentru asta singura soluție este să se extindă în univers. Cu toată viteza spre spațiu!

OBȚINEREA RESURSELOR PRIN COLONIZAREA COSMOSULUI

Cât de mare e zestrea noastră cosmică? Mai precis, ce limite superioare impune fizica asupra cantității de materie pe care viața ar putea-o în cele din urmă folosi? Desigur, zestrea noastră cosmică e uluitor de mare, dar, în mod exact, cât de mare? Tabelul 6.2 prezintă câteva cifre semnificative. Planeta noastră este acum 99,999999% moartă, în sensul că acest procentaj din materia ei nu face parte din biosferă și nu face aproape nimic util pentru viață, în afara faptului că oferă atracție gravitațională și un câmp magnetic. Există deci în principiu posibilitatea ca într-o bună zi să fie folosită de o sută de milioane de ori mai multă materie în beneficiul vieții. Dacă am putea utiliza optim materia din sistemul nostru solar (inclusiv Soarele), ne-am descurca de încă o sută de milioane de ori mai bine. Colonizarea galaxiei ne-ar spori resursele de încă un bilion de ori.

Cât de departe putem merge?

Vă închipuiți pesemne că putem obține resurse nelimitate colonizând oricâte galaxii vrem, dacă suntem suficient de perseverenți,

* Am vorbit doar despre materia alcătuită din atomi. Există de vreo șase ori mai multă materie întunecată, dar e foarte greu de detectat și captat, așa încât rămâne de văzut dacă viața viitoare o va putea utiliza. (N. a.) [Tot ce știm despre materia întunecată este că, din motive de ordin astrofizic, ar trebui să existe, dar deocamdată n-a fost observată direct, și nici nu se cunoaște compoziția ei; a nu se confunda cu *energia întunecată* (responsabilă pentru expansiunea accelerată a universului) despre care va fi vorba mai jos – n. l.]

Regiune	Particule
Biosfera noastră	10^{43}
Planeta noastră	10^{51}
Sistemul nostru solar	10^{57}
Galaxia noastră	10^{69}
Zona accesibilă călătorind cu jumătate din viteza luminii	10^{75}
Zona accesibilă călătorind cu viteza luminii	10^{76}
Universul nostru	10^{78}

Tabelul 6.2 Numărul aproximativ al particulelor de materie (protoni și neutroni) pe care viața viitoare ar putea spera să le folosească.

Însă nu asta ne sugerează cosmologia modernă. Da, spațiul însuși s-ar putea să fie infinit, conținând infinit de multe galaxii, stele și planete – și într-adevăr asta prezic cele mai simple versiuni ale *inflației*, paradigma științifică actuală care explică big bang-ul de acum 13,8 miliarde de ani. Dar chiar dacă există infinit de multe galaxii, vedem doar vreo 200 de miliarde și putem coloniza cel mult 10 miliarde de galaxii.

Ceea ce ne limitează este viteza luminii: un an-lumină (aproximativ 10 bilioane de kilometri) pe an. Figura 6.6 prezintă partea din spațiu de unde lumina a ajuns la noi în cursul celor 13,8 miliarde de ani scurși de la big bang, o regiune sferică numită „universul nostru observabil” sau, mai simplu, „universul nostru”. Chiar dacă spațiul e infinit, universul nostru e finit, conținând „doar” aproximativ 10^{78} atomi. Apoi, circa 98% din universul nostru „poate fi văzut, dar nu și atins”, în sensul că-l putem vedea, dar niciodată nu vom ajunge la el, chiar dacă am călători la nesfârșit cu viteza luminii. De ce oare? La urma urmei, limita până la care putem vedea e dată de faptul că universul nostru nu are o vechime infinită, așa încât lumina de la mare distanță n-a avut încă timp să ajungă la noi. N-am putea așadar călători până la galaxii oricât de îndepărtate dacă nu ne impunem vreo limită pentru durata călătoriei?

Primul impediment ține de faptul că universul nostru se extinde, ceea ce înseamnă că aproape toate galaxiile se îndepărtează

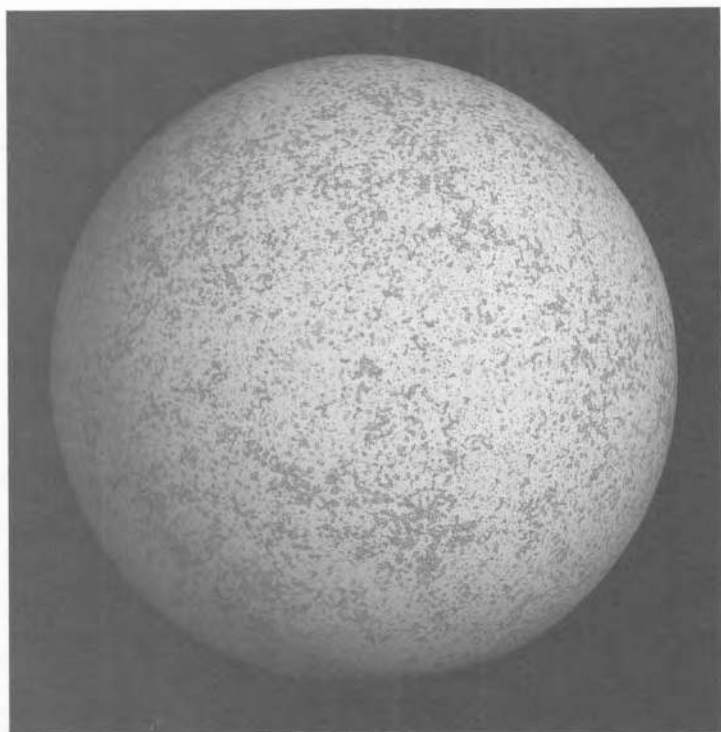


Figura 6.6 Universul nostru – regiunea sferică din spațiu de unde lumina a avut timp să ajungă la noi (ne aflăm în centru) în cursul celor 13,8 miliarde de ani scurși de la big bang. Structurile de pe suprafață sunt „fotografii de bebeluș” ale universului obținute de satelitul Planck și arată că, pe când avea vârsta de doar 400 000 de ani, era alcătuit din plasmă fierbinte, cu o temperatură aproape la fel de ridicată ca cea de la suprafața Soarelui. Probabil că spațiul se continuă dincolo de această regiune, iar în fiecare an ni se dezvoltă materie nouă.

de noi, așa încât colonizarea galaxiilor aflate la mare distanță e un fel de joc de-a prinselea. Al doilea impediment ține de faptul că această expansiune cosmică e accelerată, datorită misterioasei energii întunecate care contribuie cu circa 70% la energia totală a universului nostru. Pentru a înțelege ce problemă creează asta, imaginează-ți că ajungi pe peronul gării și vezi că trenul tău accelerează încet îndepărtându-se de tine, dar o ușă a rămas deschisă.

Dacă ești rapid și temerar, poți oare prinde trenul? În cele din urmă va merge mai repede decât poți tu alerga, deci răspunsul depinde de distanța la care te afli inițial față de tren: dacă e mai mare decât o anumită valoare critică, n-ai să poți prinde trenul niciodată. Suntem în aceeași situație încercând să ajungem din urmă acele galaxii de la mare distanță care se îndepărtează accelerat: chiar dacă am călători cu viteza luminii, toate galaxiile aflate la o distanță mai mare de 17 miliarde de ani-lumină rămân pe veci de neatins – iar ele reprezintă peste 98% din galaxiile universului nostru.

Dar stați o clipă: nu spune oare teoria relativității speciale a lui Einstein că nimic nu poate călători mai rapid decât lumina? Atunci cum de întrec galaxiile ceva ce călătorește cu viteza luminii? Răspunsul este că teoria relativității speciale este înlocuită de teoria relativității generale a lui Einstein, care e mai permisivă în privința limitării vitezei: nimic nu poate călători mai repede decât lumina *prin spațiu*, dar spațiul e liber să extindă cât de repede vrea. Einstein ne-a oferit și o modalitate de a vizualiza aceste limite ale vitezei, considerând timpul ca a patra dimensiune a spațiului-timp (vezi figura 6.7, în care am păstrat doar trei dimensiuni, omițând una dintre cele trei dimensiuni spațiale). Dacă spațiul nu s-ar extinde, razele de lumină ar fi linii drepte înclinate la 45° prin spațiu-timp, așa încât regiunile pe care le-am putea vedea și la care am putea ajunge de aici ar fi conuri. Conul nostru luminos din trecut ar fi trunchiat de big bang-ul din urmă cu 13,8 miliarde de ani, iar conul nostru luminos din viitor s-ar extinde pentru totdeauna, dându-ne acces la o zestre cosmică nelimitată. În schimb, imaginea din centrul figurii arată că un univers în expansiune cu energie întunecată (așa cum pare să fie universul nostru) deformează conurile luminoase făcându-le să semene cu un pahar de șampanie, și limitând pe veci la circa 10 miliarde numărul galaxiilor pe care le putem coloniza.

Dacă această limită vă face să simțiți claustrofobie cosmică, vreau să vă ridic moralul spunându-vă că în calculele mele am presupus că energia întunecată rămâne constantă de-a lungul timpului, ceea ce este în acord cu cele mai recente măsurători. Nu

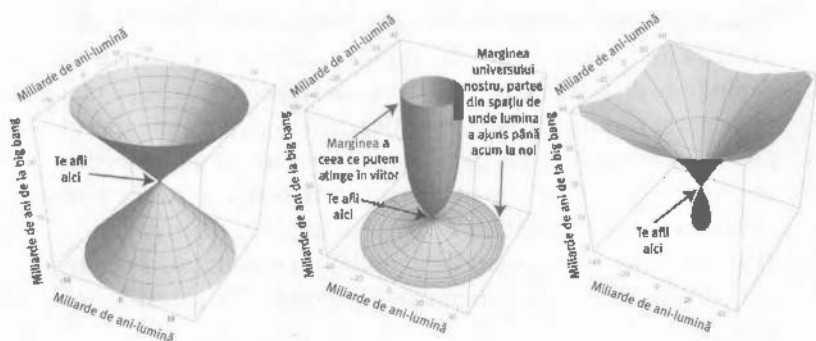


Figura 6.7 Într-o diagramă spațio-temporală, un eveniment este un punct ale cărei poziții pe orizontală și pe verticală reprezintă unde și, respectiv, când are loc. Dacă spațiul nu se extinde (imaginea din stânga), două conuri delimitează regiunea din spațiu-timp care pe noi, aflați pe Pământ (vârful conurilor), ne poate influența (conul de jos), și cea pe care o putem influența (conul de sus). Lucrurile devin mai interesante atunci când spațiul se extinde (imaginile din centru și dreapta). Conform modelului standard din cosmologie, putem vedea și atinge doar o parte finită a spațiului-timp, chiar dacă spațiul e infinit. În imaginea din mijloc, semănând cu un pahar de șampanie, folosim coordonate care ascund expansiunea spațiului, așa încât mișcările galaxiilor în decursul timpului corespund dreptelor verticale. Din punctul nostru actual de observație, la 13,8 miliarde de ani după big bang, au avut timp să ajungă la noi doar razele de lumină de la baza paharului de șampanie, și, chiar dacă am călători cu viteza luminii, n-am putea niciodată ajunge la regiunile din afara părții superioare a paharului, care conține aproximativ 10 miliarde de galaxii. În imaginea din dreapta, care seamănă cu o picătură de apă aflată sub o floare, folosim coordonatele obișnuite în care se vede expansiunea spațiului. Aceasta deformează baza paharului făcând-o să se mene cu o picătură, pentru că regiunile de la marginea a ceea ce putem vedea erau la început foarte aproape una de alta.

știm însă deocamdată ce este de fapt energia întunecată, așa că există o mică speranță ca aceasta să dispară în cele din urmă (la fel cum a dispărut substanța asemănătoare energiei întunecate a cărei existență a fost postulată pentru a explica inflația cosmică), iar dacă așa se va întâmpla, accelerația va fi înlocuită de *decelerație*, permițând în principiu viitoarelor forme de viață să continue să colonizeze mereu alte galaxii atâta vreme cât vor dăinui.

Cât de repede ne putem deplasa?

Am văzut mai sus câte galaxii poate coloniza o civilizație dacă se răspândește în toate direcțiile cu viteza luminii. Teoria relativității ne spune că e imposibil să trimiți rachete prin spațiu cu viteza luminii, fiindcă ai avea nevoie de energie infinită – prin urmare, cât de repede se pot deplasa practic rachetele?*

Când a fost lansată spre Pluto în 2006, racheta NASA New Horizons a bătut recordul de viteză, ajungând la 45 km/s, iar sonda spațială Parker urma să atingă în 2018 o viteză de patru ori mai mare în apropierea Soarelui, însă chiar și această valoare reprezintă doar 0,1% din viteza luminii. Proiectarea unor rachete mai rapide a captivat unele dintre mințile strălucite ale secolului trecut, și există o literatură bogată și fascinantă pe această temă. De ce e atât de greu să te deplasezi mai repede? Există două mari probleme: rachetele convenționale cheltuiesc cea mai mare parte a combustibilului pur și simplu pentru a accelera combustibilul pe care îl transportă, iar combustibilul de azi e îngrozitor de inefficient – fracțiunea din masa lui transformată în energie nu e mult mai mare decât cea fracțiune de 0,0000005% pentru benzină din tabelul 6.1. O îmbunătățire evidentă ar fi trecerea la un combustibil mai eficient. De exemplu, o echipă din care făcea parte și Freeman Dyson a lucrat la Proiectul Orion al NASA, în care, prin detonarea a 300 000 de bombe nucleare în cursul a zece zile, o navă suficient de mare pentru a transporta oameni într-o călătorie de un secol spre alt sistem solar să fie accelerată până la 3% din viteza luminii.⁵ Altă idee presupunea folosirea antimateriei drept combustibil, fiindcă prin combinarea acesteia cu materia obișnuită se eliberează energie cu o eficiență de aproape 100%.

Altă propunere este de a construi o rachetă care nu-și transportă combustibilul. De exemplu, spațiul interstelar nu e complet vid,

* Un calcul simplu arată că, dacă civilizația se răspândește prin spațiu nu cu viteza luminii c , ci o cu o viteză v mai mică, numărul galaxiilor colonizate scade cu un factor $(v/c)^3$. Asta înseamnă că o civilizație mai leneșă, care se răspândește cu o viteză de 10 ori mai mică decât cea a luminii, va coloniza de 1 000 de ori mai puține galaxii. (*N. a.*)

ci conține cantități mici de ioni de hidrogen (atomi de hidrogen care și-au pierdut electronul, adică protoni). Asta i-a sugerat în 1960 fizicianului Robert Bussard ideea a ceea ce numim acum *colector Bussard*: în traiectoria ei, racheta culege acești ioni și îi folosește drept combustibil într-un reactor de fuziune aflat la bord. Cercetări recente arată că punerea în practică a ideii e îndoielnică, dar există o altă propunere de navigare spațială fără a transporta combustibil, care pare să fie realistă pentru o civilizație cu tehnologie avansată: navigarea cu ajutorul laserului.

Figura 6.8 ilustrează un proiect de rachetă propulsată prin laser, conceput în 1984 de Robert Forward, același care a propus construirea unei sfere Dyson monolitice staționare. La fel cum moleculele de aer ricoșând pe vela unei ambarcațiuni o împing, particulele de lumină (fotonii) ricoșând pe o oglindă împing racheta. Orientând fasciculul unui laser uriaș alimentat cu energie solară spre o imensă velă foarte ușoară atașată de navă, putem folosi energia Soarelui pentru a accelera racheta până la viteze mari. Dar cum o oprim? Am aflat răspunsul din strălucitul articol al lui Forward: după cum se vede în figura 6.8, inelul exterior al velei se detașează și e deplasat în fața navei, reflectând fasciculul laser pentru a decelera nava și vela mai mică.⁶ Forward a calculat că în felul acesta am putea face călătoria de patru ani-lumină până la sistemul solar α Centauri în doar 40 de ani. Ajunși acolo, am construi un alt laser gigant și ne-am continua din stea în stea călătoria prin Calea Lactee.

De ce să ne oprim însă aici? În 1964, astronomul sovietic Nikolai Kardașev a propus clasificarea civilizațiilor după cantitatea de energie pe care o pot folosi. Exploatarea energiei unei planete, a unei stele (cu o sferă Dyson, de pildă) și a unei galaxii corespunde civilizațiilor de Tipul I, Tipul II și, respectiv, Tipul III pe scara lui Kardașev. S-a propus ulterior ca Tipul IV să corespundă exploatarea întregului nostru univers accesibil. Între timp au apărut o veste bună și o veste proastă pentru formele de viață ambițioase. Vestea proastă este că există energia întunecată, care, după cum am văzut, pare să ne limiteze raza de acțiune. Vestea bună este că

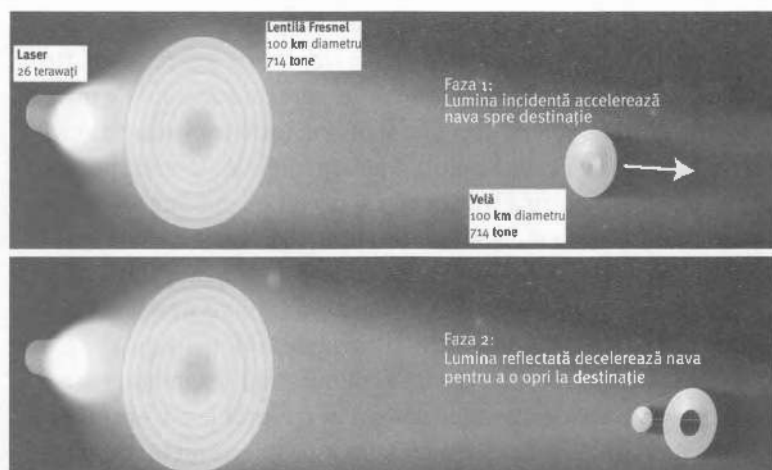


Figura 6.8 Proiectul lui Robert Forward pentru navigare cu ajutorul laserului spre steaua α Centauri, aflată la distanța de patru ani-lumină. Inițial, un laser puternic din sistemul nostru solar accelerează nava aplicând presiunea radiației asupra velei. Pentru a frâna înainte de a ajunge la destinație, partea exterioră a velei e detașată și reflectă fasciculul laser spre navă.

inteligenta artificială a făcut progrese spectaculoase. Până și vizionari optimiști precum Carl Sagan credeau că sunt șanse mici ca oamenii să ajungă în alte galaxii, dat fiind că astronauții ar muri în prima sută de ani a unei călătorii care ar dura milioane de ani, chiar dacă viteza s-ar apropia de cea a luminii. Refuzând să cedeze, ei se gândeau la congelarea astronauților pentru a le prelungi viața, la încetinirea îmbătrânirii prin deplasarea cu viteză mare sau la trimiterea unei comunități care ar călători vreme de zeci de mii de generații – mai mult decât a supraviețuit până acum specia umană.

Posibilitatea superinteligentei schimbă complet datele problemei, făcând să crească optimismul amatorilor de călătorii intergalactice. Înlăturând nevoia de a transporta sisteme masive care să susțină viața de tip uman și adăugând tehnologia inventată de IA, colonizarea intergalactică încetează să mai fie un vis. Navigarea cu ajutorul laserului devine mult mai ieftină dacă racheta nu conține decât o „sondă-germene”; un robot capabil să se așeze pe un

asteroid sau pe o planetă din sistemul solar vizat și să construiască o nouă civilizație. Nu e nevoie nici să transporte instrucțiuni, trebuie doar să construiască o antenă receptoare suficient de mare pentru a primi planuri detaliate și instrucțiuni transmise de civilizația-mamă cu viteza luminii. Apoi, își folosește laserele pe care le-a construit pentru a trimite noi sonde-germene care să colonizeze simultan sistemele solare ale galaxiei. Chiar și vastele spații intergalactice conțin destule stele (care și-au părăsit galaxiile) ce pot fi folosite pentru escale, făcând astfel posibilă o strategie a „saltului de la o insula la alta” pentru navigarea intergalactică.

Odată ce IA superinteligentă a colonizat un sistem solar sau o galaxie, aducerea oamenilor acolo n-ar fi o problemă – presupunând că oamenii au reușit să facă din asta obiectivul IA. Întreaga informație necesară despre oameni poate fi transmisă cu viteza luminii, după care IA ar asambla cuarci și electroni în oamenii doriți. Aceasta s-ar putea face fie folosind o tehnologie destul de primitivă, transmitând pur și simplu cei doi gigabyți de informație necesari pentru specificarea ADN-ului unei persoane și incubând apoi un prunc care să fie crescut de IA, fie prin nanoasamblarea de către IA din cuarci și electroni a unui adult care ar avea scanate toate amintirile originalului de pe Pământ.

Asta înseamnă că, presupunând că are loc o explozie a inteligenței, întrebarea-cheie nu este dacă e posibilă colonizarea intergalactică, ci cât de repede se poate ea desfășura. Cum toate ideile prezentate aici provin de la oameni, ele trebuie privite doar ca limite inferioare ale vitezei de expansiune a vieții; viața superinteligentă ambițioasă s-ar descurca pesemne mult mai bine și ar fi puternic stimulată să împingă și mai departe limitele, căci în cursa contra timpului și a energiei întunecate fiecare creștere de 1% a vitezei medii de colonizare înseamnă cu 3% mai multe galaxii colonizate.

De exemplu, dacă durează 20 de ani să parcurgi 10 ani-lumină până la următorul sistem solar folosind metoda laser-velă, iar alți 10 ani ca să-l colonizezi și să construiești acolo noi lasere și sonde-germene, regiunea colonizată din spațiu va fi o sferă

dilatându-se în toate direcțiile în medie cu o treime din viteza luminii. Într-o analiză aprofundată din 2014 privind civilizațiile care se extind la scară cosmică, fizicianul american Jay Olson propunea o alternativă la „saltul de pe o insulă pe alta”, implicând două tipuri de sonde: *sonde-germene* și *expandoare*.⁷ Sondele-germene ar încetini, s-ar așeza pe destinațiile lor și ar însămânța acolo viața. Expandoarele însă nu s-ar opri niciodată: ar aduna materie în cursul zborului, folosind pesemne o variantă perfecționată a tehnologiei collectorului, și ar folosi materia drept combustibil și pentru a construi sonde-germene și copii ale lor însele. Această flotă auto-reproducătoare de expandoare ar accelera ușor pentru a păstra mereu o viteză constantă (să zicem, jumătate din viteza luminii) în raport cu galaxiile învecinate, și s-ar reproduce suficient de des pentru ca flota să formeze o coajă sferică în expansiune, cu un număr constant de expandoare pe aria cojii.

În fine, dar nu în ultimul rând, există ingenioasa abordare „Ave Maria” pentru o expansiune și mai rapidă: folosind trucul „spa-mului cosmic” al lui Hans Moravec din capitolul 4. Trimițând un mesaj care păcălește civilizațiile naive proaspăt evolute pentru a construi o mașinărie superinteligentă care le jefuiește, o civilizație se poate extinde practic cu viteza luminii, viteza cu care seducătorul ei cântec de sirenă se răspândește prin cosmos. Din moment ce aceasta ar putea fi *singura* soluție pentru civilizațiile avansate de a ajunge la majoritatea galaxiilor din conul luminos al viitorului, iar ele n-ar prea avea motive să n-o aplice, ar trebui să fim foarte precauți cu mesajele primite de la extratereștri! În cartea lui Carl Sagan *Contact*, noi, pământenii, folosim planuri primite de la extratereștri pentru a construi o mașinărie pe care n-o înțelegem – nu recomand să facem asta...

În concluzie, cred că majoritatea oamenilor de știință și a autorilor SF au abordat cu pesimism problema colonizării cosmosului, ignorând posibilitatea apariției unei superinteligente: concentrându-și atenția doar asupra călătorilor umani, ei au supraestimat dificultatea călătoriei intergalactice, iar concentrându-și atenția doar asupra tehnologiei inventate de oameni, au supraestimat timpul necesar pentru a atinge limitele fizice a ceea ce este posibil.

Dacă energia întunecată continuă să accelereze îndepărtarea galaxiilor una de alta, așa cum sugerează cele mai recente observații, vor apărea mari dificultăți pentru viitorul vieții. Chiar dacă o civilizație viitoare reușește să colonizeze un milion de galaxii, în cursul a zeci de miliarde de ani energia întunecată va fragmenta acest imperiu cosmic în mii de regiuni incapabile să comunice între ele. Dacă viața viitoare nu face nimic pentru a împiedica această fragmentare, atunci cele mai mari bastioane ale vieții care vor rezista vor fi grupări conținând vreo mie de galaxii a căror gravitație combinată e suficient de puternică pentru a echilibra energia întunecată care încearcă să le separe.

Dacă o civilizație superinteligentă vrea să rămână conectată, va trebui să întreprindă lucrări de inginerie cosmică la scară mare. Cât de multă materie va avea timp să adune în cea mai mare aglomerare a ei, înainte ca energia întunecată s-o plaseze pentru totdeauna dincolo de raza ei de acțiune? O metodă de a deplasa o stea pe distanțe mari este aceea de a împinge o a treia stea într-un sistem binar în care două stele se rotesc stabil una în jurul celeilalte. La fel ca în relațiile amoroase, introducerea unui al treilea partener poate avea un efect destabilizator, unul dintre cei trei parteneri fiind violent respins – în cazul stelelor, cu viteză mare. Dacă unii dintre parteneri sunt găuri negre, o asemenea tripletă instabilă poate fi folosită pentru a azvârli materie în afara galaxiei-gazdă. Din păcate, prin această tehnică a celor trei corpuri, fie că e aplicată stelelor, găurilor negre sau galaxiilor, s-ar părea că nu poate fi deplasată decât o infimă parte din masa civilizației pe distanțele mari necesare pentru a învinge energia întunecată.

Asta nu înseamnă însă că viața superinteligentă nu poate găsi soluții mai bune – de pildă, transformând o mare parte din masa galaxiilor îndepărtate în nave spațiale care pot călători spre aglomerarea centrală. Dacă poate fi construit un sfalerizor, pesemne că acesta ar putea fi folosit pentru a converti materia în energie care să fie trimisă spre aglomerarea centrală sub formă de lumină, iar acolo lumina să fie reconfigurată în materie sau utilizată ca atare.

Cea mai mare șansă ar fi dacă am reuși să construim găuri de vierme (*wormholes*) stabile traversabile, permițând comunicarea și călătoria aproape instantanee între cele două capete ale găurii de vierme, indiferent cât de mare e distanța dintre ele. Gaura de vierme este o „scurtătură” în spațiu-timp prin care ajungi din punctul A în punctul B fără a străbate spațiul dintre cele două puncte. Deși găurile de vierme stabile nu sunt interzise de relativitatea generală și au apărut în filme precum *Contact* și *Interstellar*, ele presupun existența unui ipotetic tip de materie cu densitate negativă, care depinde de efecte ale gravitației cuantice încă necunoscute. Cu alte cuvinte, s-ar putea dovedi că găurile de vierme utile sunt imposibile – în caz contrar însă, viața superinteligentă are toate motivele să le construiască. Găurile de vierme nu numai că revoluționează comunicarea rapidă între galaxii, dar legând de timpuriu galaxiile îndepărtate de aglomerarea centrală, găurile de vierme ar permite domeniului vieții viitoare să rămână conectat un timp nedefinit, anulând complet încercările energiei întunecate de a cenzura comunicarea. Odată ce două galaxii sunt legate printr-o gaură de vierme stabilă, ele rămân legate oricât s-ar îndepărta una de alta.

Dacă, în ciuda eforturilor de inginerie cosmică, o civilizație a viitorului conchide că e condamnată să piardă pe veci legătura cu zone ale ei, s-ar putea pur și simplu despărți de ele, urându-le succes. Dacă însă are obiective de calcul ambițioase, implicând căutarea răspunsurilor la întrebări extrem de dificile, ar putea apela la strategia „pământului pârjolit”: ar converti galaxiile îndepărtate în calculatoare masive care transformă materia și energia lor în calcule efectuate într-un ritm frenetic, cu speranța că, înainte ca energia întunecată să împingă resturile lor epuizate dincolo de orizont, acestea ar putea transmite răspunsurile lor îndelung căutate către aglomerarea centrală. Această strategie a „pământului pârjolit” ar fi potrivită mai cu seamă regiunilor atât de îndepărtate, încât se poate ajunge la ele doar prin metoda „spamului cosmic”, spre disperarea locuitorilor lor. În schimb, în aglomerarea centrală, civilizația s-ar strădui ca prin maximă conservare și eficiență să reziste cât mai mult cu putință.

Cât de mult poți dăinui?

Majoritatea persoanelor, organizațiilor și națiunilor ambițioase aspiră la longevitate. Prin urmare, dacă o civilizație ambițioasă a viitorului dezvoltă superinteligenta și vrea longevitate, cât de mult poate dăinui?

Prima analiză științifică aprofundată a viitorului nostru îndepărtat a fost efectuată de nimeni altul decât Freeman Dyson, iar tabelul 6.3 rezumă câteva dintre rezultatele sale importante. Concluzia este că, fără intervenția inteligenței, sistemele solare și galaxiile sunt treptat distruse, urmând apoi tot restul și rămân doar spațiul vid, rece și mort, cu o strălucire a radiației stingându-se în eternitate. Dar Freeman își încheie analiza pe un ton optimist: „Există motive științifice întemeiate pentru a vorbi cu seriozitate despre posibilitatea ca viața și inteligența să modeleze acest univers al nostru conform scopurilor lor.”⁸

Ce	Când
Vârsta actuală a universului	10^{10} ani
Energia întunecată împinge majoritatea galaxiilor, făcându-le inaccesibile	10^{11} ani
Ultimele stele au ars complet	10^{14} ani
Planetele se detașează de stele	10^{15} ani
Stelele se detașează de galaxii	10^{19} ani
Destabilizarea orbitelor prin radiație gravitațională	10^{20} ani
Dezintegrarea protonilor (cel mai curând)	$> 10^{34}$ ani
Evaporarea găurilor negre cu mase cât ale stelelor	10^{67} ani
Evaporarea găurilor negre supermasive	10^{91} ani
Întreaga materie se dezintegrează transformându-se în fier	10^{1500} ani
Întreaga materie formează găuri negre, care apoi se evaporă	$10^{10^{26}}$

Tabelul 6.3 Estimări privind viitorul îndepărtat aparținând (cu excepția celei de-a doua și celei de-a șaptea) lui Freeman Dyson. El a efectuat aceste calcule înainte de descoperirea energiei întunecate, care ar putea provoca anumite tipuri de „cosmocalipse” peste 10^{10} – 10^{11} ani. Protonii ar putea fi absolut stabili; în caz contrar, experimentele sugerează că vor trece peste 10^{34} ani înainte ca jumătate dintre ei să se dezintegreze.

Cred că superinteligenta ar rezolva lesne multe dintre probleme enumerate în tabelul 6.3, pentru că poate rearanja materia în ceva mai bun decât sisteme solare și galaxii. Provocări deseori discutate, cum e moartea Soarelui nostru peste câteva miliarde de ani, nu ne vor opri din drum, fiindcă până și o civilizație cu un nivel tehnologic relativ modest se poate cu ușurință muta pe stele cu masă mică, al căror timp de viață depășește 200 de miliarde de ani. Presupunând că civilizațiile superinteligente își construiesc propriile lor centrale energetice mai eficiente decât stelele, ele și-ar dori să împiedice formarea stelelor pentru a conserva energia: chiar dacă ar folosi o sferă Dyson pentru a colecta întreaga energie produsă de o stea în timpul ei de viață principal (obținând aproximativ 0,1% din energia totală înmagazinată în stea), s-ar putea să nu recupereze prea mult din restul de 99,9% atunci când o stea foarte masivă moare. O stea masivă sfârșește printr-o supernovă, o explozie în care majoritatea energiei e disipată sub formă de neutrini, particule greu de detectat, iar la stelele foarte masive o mare cantitate de masă e irosită în formarea unei găuri negre din care energia se scurge în 10^{67} ani.

Câtă vreme viața superinteligentă nu și-a epuizat materia/energia, ea își poate menține habitatul în starea pe care și-o dorește. Peseamne că ar putea descoperi un mijloc de a împiedica protonii să se dezintegreze folosind așa-numitul efect *watched-pot** din mecanica cuantică, prin care procesul de dezintegrare e încetinit dacă faci observații în mod regulat. Există însă riscul ca peste 10–100 de miliarde de ani să se producă o *cosmocalipsă* care să distrugă întregul nostru univers. Descoperirea energiei întunecate și cercetări din teoria corzilor conduc la noi scenarii ale cosmocalipsei pe care Freeman Dyson nu le cunoștea când și-a scris articolul.

Cum va sfârși așadar universul nostru peste miliarde de ani? Există cinci scenarii principale pentru apocalipsa cosmică, sau

* Expresia derivă din proverbul englezesc „A watched pot never boils” („Oala pe care o păzești nu fierbe”). Numit și efect Zenon cuantic, acesta se bazează pe faptul că o măsurătoare efectuată asupra unui sistem cuantic influențează starea sistemului, deci măsurători frecvent repetate ar putea „îngheța” sistemul într-o anumită stare, blocându-i evoluția. (N. t.)

cosmocalipsa, ilustrate în figura 6.9: *marele îngheț* (*big chill*), *marea implozie* (*big crunch*), *marea ruptură* (*big rip*), *marea plesnitură* (*big snap*) și *bule mortale* (*death bubbles*). Universul nostru se extinde de circa 14 miliarde de ani. Marele îngheț presupune ca universul să se extindă pe veci, diluând cosmosul într-un spațiu rece, întunecos și, în cele din urmă, mort; acesta era considerat rezultatul final cel mai probabil atunci când Freeman și-a scris articolul. Îl consider alegerea lui T.S. Eliot: „Așa sfârșește lumea / Nu în explozie, ci în scâncet.” Dacă, asemenea lui Robert Frost, ați vrea ca lumea să sfârșească în flăcări, nu în gheață, rugați-vă să aibă loc marea implozie, caz în care expansiunea cosmică e în cele din urmă inversată, iar totul se prăbușește într-un colaps catastrofal, ca un big bang derulat în sens invers. Marea ruptură seamănă oarecum cu marele îngheț, galaxiile, planetele și atomii se dezmembrează într-un final grandios peste un timp finit. Pe care dintre aceste trei scenarii ai paria? Asta depinde de ce se întâmplă cu energia întunecată, care reprezintă aproximativ 70% din masa universului nostru, pe măsură ce spațiul continuă să se extindă. Poate fi oricare dintre scenariile îngheț, implozie sau ruptură, după cum energia întunecată rămâne neschimbată, se diluează până la o densitate negativă sau, respectiv, se antidiluează până la o densitate mai mare. Fiindcă deocamdată n-avem nici un indiciu privind energia întunecată, iată cum aş paria eu: 40% pe marele îngheț, 9% pe marea implozie și 1% pe marea ruptură.

Dar ce-aș face cu restul de 50% din banii mei? Îi economisesc pentru o posibilitate care nu e nici una dintre acestea trei, fiindcă ar trebui să recunoaștem că există lucruri fundamentale pe care deocamdată nu le înțelegem. Natura spațiului, de pildă. Sfârșitul prin îngheț, implozie sau ruptură presupune că spațiul însuși e stabil și poate fi întins oricât de mult. Pe vremuri, credeam că spațiul e doar scena inertă pe care se desfășoară drama cosmică. Apoi Einstein ne-a învățat că spațiul este unul dintre actorii principali: se poate curba pentru a da naștere găurilor negre, se poate undui sub formă de unde gravitaționale și se poate întinde ca univers în expansiune. Pesemne că poate chiar îngheța într-o altă fază,

CUM VA SFÂRȘI UNIVERSUL NOSTRU?

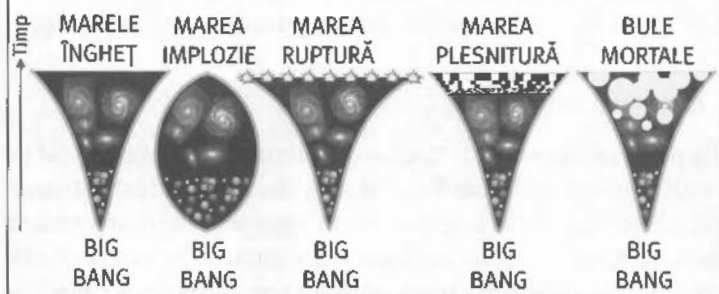


Figura 6.9 Știm că universul nostru s-a născut dintr-un big bang fierbinte cu 14 miliarde de ani în urmă, s-a extins și s-a răcit, a agregat particulele în atomi, stele și galaxii. Dar nu știm care e soarta lui. Printre scenariile propuse se numără un mare îngheț (expansiune eternă), o mare implozie (colaps), o mare ruptură (o rată infinită de expansiune care dezmembrează totul), o mare plesnitură (textura spațiului dezvăluind o natură granulară letală atunci când e întinsă prea mult) și bule mortale (spațiul „înghețând” în bule letale care se extind cu viteza luminii).

oarecum asemănător apei, iar bulele mortale ale noii faze care se extind rapid devin un nou candidat pentru cosmocalipsă. Dacă bulele mortale sunt posibile, ele s-ar extinde probabil cu viteza luminii, la fel ca sfera în permanentă dilatare a spamului cosmic trimis de o civilizație extrem de agresivă.

Teoria lui Einstein mai spune că întinderea spațiului poate continua nelimitat, permițând universului să ajungă la un volum infinit, ca în scenariile înghețului și rupturii – ceea ce bănuiesc că nu-i adevărat. O bandă de cauciuc pare continuă, la fel ca spațiul, dar dacă o întinzi prea mult, plesnește. De ce? Fiindcă e alcătuită din atomi, iar dacă e întinsă suficient de mult, devine importantă această natură atomică granulară a benzii. Ar putea oare avea și spațiul o asemenea structură granulară la o scară prea mică pentru ca noi s-o fi observat? Cercetările de gravitație cuantică sugerează că n-are sens să vorbim despre spațiul tridimensional tradițional la scări mai mici de aproximativ 10^{-34} metri. Dacă e adevărat că spațiul nu poate fi întins la nesfârșit fără a

suferi o „mare plesnitură” catastrofală, atunci civilizațiile viitorului și-ar dori pesemne să se mute în cea mai mare regiune a spațiului care nu se extinde (un roi uriaș de galaxii) la care pot ajunge.

Cât de mult poți calcula?

După ce am văzut cât de mult *poate* dăinui viața, să vedem cât de mult *și-ar dori* să dăinuie. Pare firesc să vrei să trăiești cât mai mult, însă Freeman Dyson a venit și cu un argument cantitativ pentru această dorință: costul calculului scade atunci când calculezi mai lent, așadar în cele din urmă obții mai mult dacă încetinești lucrurile cât mai mult cu putință. Freeman chiar a calculat că, dacă universul nostru continuă pe veci să se extindă și să se răcească, e posibilă o cantitate infinită de calcule.

Lent nu înseamnă neapărat plictisitor: dacă viața viitoare trăiește într-o lume simulată, curgerea timpului percepută subiectiv nu are neapărat legătură cu ritmul glaciilor în care simularea se derulează în lumea exterioară, deci perspectivele unui calcul infinit s-ar putea traduce printr-o nemurire subiectivă pentru formele de viață simulate. Cosmologul Franck Tipler a pornit de aici pentru a lansa ideea că am putea ajunge la nemurirea subiectivă în momentele finale dinaintea marii implozii, făcând ca viteza de calcul să tindă spre infinit atunci când temperatura și densitatea cresc enorm.

Din moment ce energia întunecată pare să anuleze atât speranța lui Freeman, cât și pe cea a lui Franck de a ajunge la calculul infinit, superinteligenta viitorului ar putea prefera să-și consume rezervele de energie relativ rapid, să le transforme în calcul înainte de a se lovi de dificultăți cum sunt cele legate de orizonturile cosmice și de dezintegrarea protonului. Dacă maximizarea cantității totale de calcul este obiectivul suprem, strategia optimă va fi un compromis între prea lent (pentru a evita dificultățile pomenite mai sus) și prea rapid (cheltuind mai multă energie per calcul decât e nevoie).

Tot ce am discutat în acest capitol ne arată că centrale și calculatoare de eficiență maximă ar permite vieții superinteligente să

efectueze o cantitate extraordinară de calcule. Alimentarea creierului nostru de 13 wați vreme de o sută de ani necesită energia din aproximativ o jumătate de miligram de materie – mai puțin decât un grăunte de zahăr. Conform studiilor lui Seth Lloyd, creierul ar putea avea o eficiență energetică de un trilion (10^{15}) de ori mai mare, permițând acelui grăunte de zahăr să alimenteze o simulare a tuturor vieților oamenilor care au trăit vreodată și a vieților de încă o mie de ori mai multor oameni. Dacă toată materia din universul nostru disponibil ar fi folosită pentru a simula oameni, s-ar ajunge la peste 10^{69} vieți – sau ce-ar prefera superinteligenta IA să facă cu puterea ei de calcul. Și mai multe vieți ar fi posibile dacă simulările lor ar fi rulate mai lent.⁹ Pe de altă parte, în cartea sa *Superintelligence*, Nick Bostrom estimează că, pornind de la presupuneri mai puțin optimiste privind eficiența energetică, ar putea fi simulate 10^{58} vieți umane. Oricât am analiza și reanaliza cifrele, ele rămân uriașe, la fel ca responsabilitatea noastră de a nu irosi acest potențial al vieții viitoare. „Dacă am reprezenta întreaga fericire trăită în cursul unei asemenea vieți printr-o singură lacrimă de bucurie“, spune Bostrom, „atunci fericirea acestor suflete ar umple și reumple oceanele Pământului în fiecare secundă vreme de o sută de miliarde de miliarde de milenii. Important este să ne asigurăm că acestea sunt cu adevărat lacrimi de bucurie.“

IERARHII COSMICE

Viteza luminii limitează nu doar răspândirea vieții, ci și natura ei, impunând constrângeri drastice asupra comunicării, conștiinței și controlului. Așadar, dacă o mare parte din cosmosul nostru va căpăta în cele din urmă viață, cum va arăta această viață?

Ierarhii ale gândirii

Ți s-a întâmplat vreodată să încerci să prinzi cu mâna o muscă și să nu reușești? Musca reacționează mai repede decât tine pentru că e mai mică, așa încât informației îi ia mai puțin timp să parcurgă

distanțele dintre ochi, creier și mușchi. Acest principiu „mai mare = mai lent” se aplică nu numai în biologie, unde limita de viteză e impusă de rapiditatea cu care semnalele electrice circulă prin neuroni, dar și în viața cosmică viitoare, presupunând că într-adevăr informația nu se poate propaga mai repede decât lumina. Pentru un sistem inteligent de procesare a informației, creșterea în dimensiuni prezintă deci și dezavantaje. Pe de o parte, a fi mai mare înseamnă a conține mai multă materie, ceea ce dă posibilitatea unor gânduri mai complexe. Pe de altă parte, aceasta încetinește ritmul în care pot apărea gânduri cu adevărat globale, căci acum durează mai mult până ce informația relevantă ajunge la toate componentele.

Prin urmare, dacă viața va înghiți cosmosul nostru, ce formă va alege: simplă și rapidă, sau complexă și lentă? Presupun că va face aceeași alegere pe care a făcut-o și viața de pe Pământ: ambele! Dimensiunile locuitorilor biosferei Pământului se întind pe un spectru vast, de la balena albastră cântărind două sute de tone până la bacteria de 10^{-16} kg numită *Pelagibacter*, a cărei biomasă se crede că o depășește pe cea a tuturor peștilor din lume. Apoi, la organismele mari, complexe și lente încetineala e adesea contrabalansată de modulele lor mai mici care sunt simple și rapide. De pildă, reflexul nostru de a clipi este extrem de rapid tocmai pentru că e pus în funcțiune de un circuit mic și simplu care nu implică cea mai mare parte a creierului: dacă acea muscă greu de prins se îndreaptă din întâmplare spre ochiul tău, vei clipi într-o zecime de secundă, cu mult înainte ca informația relevantă să aibă timpul să se răspândească în creier și să te facă să devii conștient de ce se petrece. Organizând procesarea informației într-o ierarhie de module, biosfera noastră reușește să obțină și viteză, și complexitate. Noi, oamenii, folosim deja aceeași strategie ierarhică pentru a optimiza calculul paralel.

Deoarece comunicarea internă e lentă și costisitoare, mă aștept ca viața cosmică avansată din viitor să procedeze la fel, așa încât calculele să fie efectuate, pe cât se poate, local. Dacă un calcul e suficient de simplu pentru a fi efectuat cu 1 kg de computer, e contraproductiv să-l răspândești într-un computer de dimensi-

unea unei galaxii, fiindcă așteptarea ca informația să fie partajată cu viteza luminii după fiecare pas de calcul provoacă o întârziere ridicolă de circa 100 000 de ani la fiecare pas.

Ce anume din această viitoare procesare a informației va fi făcut în mod *conștient* (presupunând că asta se va întâmpla), în sensul că va implica o experiență subiectivă, e o problemă delicată și fascinantă pe care o vom trata în capitolul 8. Dacă a fi conștient presupune ca diferite părți ale sistemului să poată comunica între ele, atunci gândirea sistemelor mari e în mod necesar mai lentă. În vreme ce tu sau un supercomputer viitor de dimensiunea Pământului puteți avea multe gânduri pe secundă, o minte de dimensiunea unei galaxii poate avea doar un gând la fiecare sută de mii de ani, iar o minte cosmică atingând în dimensiune un miliard de ani-lumină ar avea răgaz doar pentru zece gânduri înainte ca energia întunecată s-o fragmenteze în părți între care nici o legătură să nu fie posibilă. Nu-i mai puțin adevărat însă că aceste câteva gânduri prețioase și trăirile care le însoțesc ar putea fi nebănuite de profunde!

Ierarhii ale controlului

Dacă gândirea însăși e organizată într-o ierarhie care se întinde pe un spectru larg de dimensiuni, ce putem spune despre putere? În capitolul 4 am văzut că entitățile inteligente se organizează spontan în ierarhii ale puterii ajunse la echilibru Nash, în care orice entitate s-ar descurca mai rău dacă și-ar modifica strategia. Pe măsură ce se perfecționează tehnologia transportului și comunicării, aceste ierarhii cresc. Dacă superinteligența se va extinde într-o bună zi la scară cosmică, cum va arăta ierarhia puterii? Va fi descentralizată sau autoritară? Cooperarea se va baza în principal pe avantajul reciproc sau pe coerciție și amenințare?

Pentru a lămuri aceste probleme, trebuie să avem în vedere atât „morcovul”, cât și „bățul”: ce stimulente există pentru colaborarea la scară cosmică, și ce amenințări ar putea fi folosite pentru a o impune?

Pe Pământ, comerțul a fost un motor tradițional al cooperării pentru că dificultatea relativă de a produce anumite lucruri variază de la un loc la altul. Dacă extragerea prin minerit a unui kilogram de argint costă de 300 de ori mai mult decât extragerea unui kilogram de cupru într-o regiune, dar numai de 100 de ori mai mult decât în alta, ambele regiuni vor avea de câștigat dacă vor schimba 200 kg de cupru pentru 1 kg de argint. Dacă o regiune posedă o tehnologie mult mai avansată decât o alta, ambele ar avea de câștigat schimbând bunuri de înaltă tehnologie contra materii prime.

Dacă însă superintelența dezvoltă o tehnologie prin care poate cu ușurință rearanja particulele în orice formă de materie, aceasta va descuraja în mare parte comerțul pe distanțe lungi. De ce să te chinui să transporti argint între două sisteme solare îndepărtate, când e mai simplu și mai rapid să transformi cuprul în argint rearanjându-i particulele? De ce să te chinui să transporti dispozitive de înaltă tehnologie între galaxii, când informațiile tehnice și materiile prime (orice materie e bună) există în ambele locuri? Bănuiesc că, într-un cosmos înțesat cu superintelență, aproape singurul bun care merită transportat pe distanțe lungi va fi *informația*. Unica excepție ar putea fi materia folosită în proiectele de inginerie cosmică – de pildă, pentru a contracara tendința destructivă a energiei întunecate de a separa civilizațiile. Spre deosebire de ce se întâmplă în comerțul uman tradițional, această materie poate fi transportată sub orice formă compactă convenabilă, eventual chiar ca fascicul de energie, din moment ce superintelența care o primește o poate rapid rearanja în orice obiecte vrea.

Dacă partajarea sau vinderea informației devine principalul motor al cooperării cosmice, ce tipuri de informații ar putea fi implicate? Orice informație dorită va deveni prețioasă dacă generarea ei necesită un efort de calcul masiv și îndelungat. De exemplu, o superintelență ar putea dori să aștepte răspunsuri la întrebări științifice complicate despre natura fizică a realității, la întrebări matematice despre teoreme și algoritmi optimi sau la întrebări ingineresti privind construirea unei tehnologii spectaculoase.

Formele de viață hedoniste și-ar putea dori distracție digitală și experiențe simulate fabuloase, iar comerțul cosmic ar putea stimula cererea pentru o formă de criptomonedă cosmică în genul bitcoin.

Asemenea posibilități de schimb ar face ca informația să nu circule doar între entități comparabile, dar și în sus și în jos, în cadrul ierarhiilor de putere – de pildă, între noduri de dimensiunea unui sistem solar și un centru galactic, sau între noduri de dimensiunea unei galaxii și un centru cosmic. Nodurile și-ar dori asta pentru a simți că fac parte dintr-un ansamblu mai mare, pentru a primi răspunsuri și tehnologii la care, singure, n-au acces, și pentru a se apăra de amenințări externe. Ele ar putea fi atrase și de perspectiva unei cvasi-nemuriri prin copiere: la fel cum mulți oameni se consolează cu credința că sufletul lor va supraviețui după ce trupul fizic moare, o LA avansată s-ar mulțumi ca mintea și cunoștințele ei să supraviețuiască într-un supercalculator din centru, după ce hardware-ul ei original și-a epuizat rezervele de energie.

Reciproc, centrul ar vrea ca nodurile lui să-l ajute în sarcini computaționale masive pe termen lung, de rezultatele cărora nu are urgent nevoie, așa încât merită să aștepte răspunsul mii sau milioane de ani. Așa cum am văzut mai sus, centrul și-ar putea dori și ca nodurile să-l ajute la proiecte ingineresti de anvergură, cum ar fi contracararea energiei întunecate deplasând întregi concentrații de masă galactică. Dacă se va dovedi că găurile de vierme traversabile există și pot fi construite, atunci principala prioritate a centrului va fi pesemne construirea unei rețele de găuri de vierme pentru a se opune energiei întunecate și a păstra imperiul conectat pe termen nedefinit. Problema aprig discutată a obiectivelor supreme pe care le-ar putea avea o superintelență va fi analizată în capitolul 7.

Controlul cu ajutorul „bățului“

Imperiile terestre își constrâng de regulă subordonații să coopereze apelând atât la „morcov“, cât și la „băț“. Supușii Imperiului Roman prețuiau tehnologia, infrastructura și protecția care li se ofereau în schimbul cooperării lor, dar se și temeau de repercusiunile

inevitabile în caz de revoltă sau neplata dărilor. Datorită timpului lung necesar pentru a trimite trupe de la Roma în provinciile îndepărtate, intimidarea cădea în bună măsură în sarcina trupelor locale și a înalților funcționari loiali, care aveau misiunea de a aplica imediat pedepse. Un centru superintelligent ar putea folosi strategia asemănătoare de a desfășura o rețea de gărzi loiale pe întinsul imperiului cosmic. Cum supușii superintelligenți pot fi greu de controlat, cea mai simplă strategie viabilă ar fi folosirea unor gărzi IA programate să fie 100% loiale în virtutea faptului că sunt relativ stupide, urmărind doar ca toate regulile să fie respectate, în caz contrar activând automat o armă a apocalipsei.

Să presupunem, de exemplu, că centrul IA plasează o stea pitică albă în apropierea unei civilizații de dimensiunea unui sistem solar pe care vrea s-o controleze. Pitica albă este ceea ce rămâne după ce o stea de masă medie își epuizează combustibilul. Alcătuită în mare măsură din carbon, seamănă cu un diamant uriaș, și e atât de compactă încât poate cântări mai mult decât Soarele, deși e mai mică decât Pământul. Fizicianul indian Subrahmanyan Chandrasekhar a demonstrat că, dacă îi adaugi masă până depășește *limita Chandrasekhar*, de vreo 1,4 ori masa Soarelui nostru, va exploda într-un cataclism termonuclear numit supernovă de tipul 1A. Dacă IA centrală a aranjat în mod cinic ca această pitică albă să fie foarte aproape de limita Chandrasekhar, gardianul IA ar fi eficient chiar și dacă ar fi extrem de stupid (în bună parte, tocmai datorită stupidității): ar putea fi programat să verifice numai dacă civilizația subjugată și-a predat cota lunară de bitcoini, demonstrații matematice sau orice alte taxe erau stipulate, iar în caz contrar să arunce suficientă masă în pitica albă pentru a declanșa supernova, distrugând complet întreaga regiune.

Civilizații de dimensiunea unei galaxii ar putea fi controlate în mod asemănător, plasând multe obiecte compacte pe orbite apropiate de gaura neagră uriașă din centrul galaxiei, și amenințând să transforme aceste mase în gaz – de pildă, ciocnindu-le. Acest gaz ar începe apoi să hrănească gaura neagră, transformând-o într-un quasar puternic care ar face ca o mare parte din galaxie să devină nelocuibilă.

În concluzie, există motive puternice pentru ca viața din viitor să coopereze la distanțe cosmice, dar nu putem spune dacă această cooperare se va baza în principal pe avantaje reciproce sau pe amenințări brutale – limitele impuse de fizică par să permită ambele scenarii, așa încât rezultatul va depinde de obiectivele și valorile predominante. Vom discuta despre capacitatea noastră de a influența aceste obiective și valori ale vieții viitoare în capitolul 7.

Când civilizațiile se confruntă

Am analizat până acum doar scenarii în care viața se extinde în cosmos pornind de la o unică explozie a inteligenței. Ce se întâmplă însă dacă viața evoluează independent în mai multe locuri, iar două civilizații în expansiune se întâlnesc?

Pentru un sistem solar oarecare, probabilitatea ca viața să evolueze pe una dintre planetele lui, să dezvolte o tehnologie avansată și să se extindă în spațiu pare să fie nenulă, odată ce viața tehnologică a evoluat aici, în sistemul nostru solar, iar legile fizicii nu interzic colonizarea cosmosului. Dacă spațiul e suficient de vast (teoria inflației cosmologice sugerează că e imens sau infinit), atunci există pesemne numeroase civilizații care se extind, după cum e ilustrat în figura 6.10. Articolul lui Jay Olson menționat mai sus conține o elegantă analiză a unor asemenea biosfere cosmice în expansiune, iar Toby Ord, împreună cu colegii lui de la Institutul pentru Viitorul Omenirii, a efectuat o analiză similară. Privite în cele trei dimensiuni spațiale, aceste biosfere cosmice sunt practic sfere, dacă presupunem că civilizațiile se extind cu aceeași viteză în toate direcțiile. În spațiu-timp, biosferele arată ca partea superioară a „paharului de șampanie” din figura 6.7, fiindcă energia întunecată limitează numărul galaxiilor la care fiecare civilizație poate ajunge.

Dacă distanța dintre civilizațiile colonizatoare vecine e mult mai mare decât le permite energia întunecată să se extindă, ele nu vor intra niciodată în contact, și nici măcar nu vor afla una de existența celeilalte, așa încât fiecare va simți că e singură în cosmos.

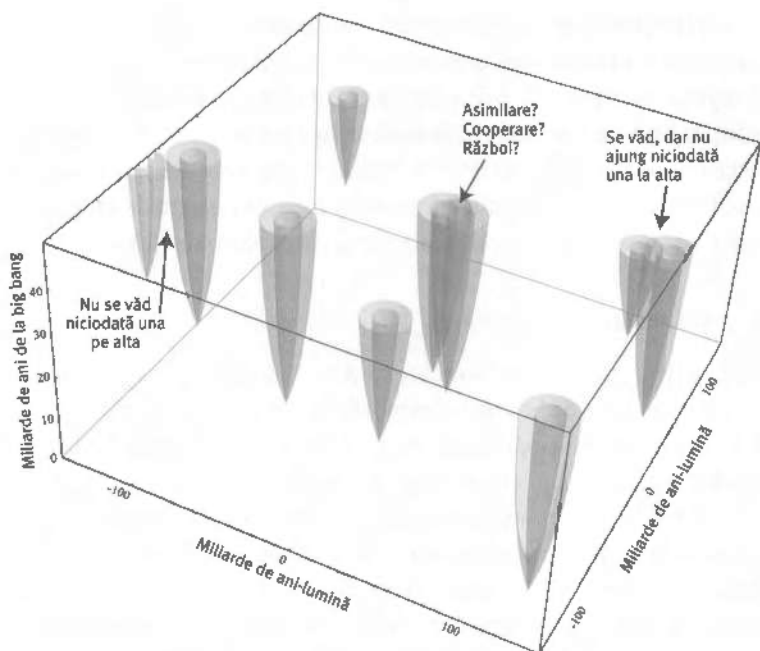


Figura 6.10 Dacă viața evoluează independent în mai multe puncte din spațiu-timp (locuri și momente) și începe să colonizeze spațiul, atunci spațiul va conține o rețea de biosfere cosmice în expansiune, fiecare semănând cu partea de sus a „paharului de șampanie” din figura 6.7. Capătul de jos al fiecărei biosfere reprezintă locul și momentul în care a început colonizarea. „Paharele de șampanie” opac și translucid corespund colonizării cu viteză de 50% și, respectiv, 100% din viteza luminii, iar suprapunerile indică unde se întâlnesc civilizațiile independente.

Dacă însă cosmosul nostru e atât de fecund încât distanțele dintre civilizațiile vecine să fie mai mici, atunci unele civilizații se vor suprapune în cele din urmă. Ce se va întâmpla în aceste regiuni de suprapunere? Va fi cooperare, competiție sau război?

Europenii au putut cuceri Africa și Americile fiindcă aveau o tehnologie superioară. Dimpotrivă, e plauzibil ca, mult înainte de întâlnirea a două civilizații superinteligente, tehnologiile lor să se fi stabilizat la același nivel, limitat doar de legile fizicii. Prin urmare, pare puțin probabil ca o superintelență s-o cucerească

lesne pe cealaltă, chiar dacă asta și-ar dori. Apoi, dacă obiectivele lor au evoluat aproximativ în aceeași direcție, nu prea ar avea motive să-și dorească cucerirea sau războiul. De pildă, dacă ambele ar încerca să demonstreze cât mai multe teoreme frumoase și să inventeze algoritmi cât mai ingenioși, ele își pot pur și simplu împărtăși descoperirile în beneficiul ambelor. În definitiv, informația e foarte diferită de resursele pentru care oamenii se luptă de obicei, odată ce o poți dăruia, și în același timp păstra.

Unele civilizații în expansiune ar putea avea obiective esențialmente inflexibile, ca acelea ale unui cult fundamentalist sau ale unui virus care se răspândește. E însă de asemenea plauzibil ca unele civilizații avansate să semene mai curând cu oamenii lipsiți de prejudecăți – gata să-și ajusteze obiectivele când li se prezintă argumente suficient de convingătoare. Dacă două dintre acestea se întâlnesc, va fi o confruntare nu de arme, ci de idei, în care triumfă cea mai persuasivă, iar obiectivele ei se răspândesc cu viteza luminii în întreaga regiune controlată de cealaltă civilizație. Asimilarea vecinilor e o strategie de expansiune mai rapidă decât colonizarea, fiindcă sfera ta de influență se poate extinde cu viteza cu care se propagă ideile (viteza luminii, folosind telecomunicația), în vreme ce colonizarea fizică avansează inevitabil mai lent. Această asimilare nu va fi forțată, ca aceea aplicată de Borg din *Star Trek*, ci voluntară, bazată pe superioritatea persuasivă a ideilor, cel asimilat având de câștigat.

Am văzut că universul viitor poate conține două tipuri de bule în expansiune rapidă: civilizații care se extind și acele bule mortale care se extind cu viteza luminii și fac ca spațiul să devină de nelocuit, distrugând toate particulele elementare. O civilizație ambițioasă poate deci întâlni trei tipuri de regiuni: zone nelocuite, bule ale vieții și bule ale morții. Dacă se teme de civilizații rivale necooperante, are toate motivele să se lanseze într-o „acaparare a terenurilor” și să colonizeze zonele nelocuite înainte s-o facă rivalii. Are însă aceeași pornire expansionistă chiar și dacă nu există alte civilizații, pur și simplu pentru a dobândi resurse înainte ca energia întunecată să le facă să devină de neatins. Ca

atare, a da peste o altă civilizație în expansiune poate fi un lucru mai bun sau mai rău decât a da peste un spațiu nelocuit, în funcție de măsura în care vecinul e cooperant și lipsit de prejudecăți. Oricum, e mai bine să te ciocnești de o civilizație expansionistă (chiar și de una care încearcă să te transforme în agrafe de birou) decât să te ciocnești de o bulă mortală, care va continua să se extindă cu viteza luminii, indiferent dacă tu încerci să te lupți sau să porți o discuție rezonabilă cu ea. Singura protecție împotriva bulelor mortale ne-o poate oferi energia întunecată, care le împiedică pe cele aflate la mare distanță să ajungă la noi. Prin urmare, dacă bulele mortale sunt într-adevăr răspândite prin univers, energia întunecată este de fapt prietenul, nu dușmanul nostru.

Suntem oare singuri?

Mulți oameni sunt convinși că există viață avansată în numeroase locuri din univers, așa încât extincția omenirii n-ar însemna mare lucru din perspectivă cosmică. În fond, de ce să ne îngrijorăm că ne-am putea autodistrage, dacă o civilizație cu obiective nobile gen *Star Trek* și-ar face curând apariția și ar reînsămânța cu viață sistemul nostru solar? Cred că această ipoteză *Star Trek* e periculoasă, fiindcă ne dă o amăgitoare senzație de securitate, iar civilizația noastră riscă să devină indiferentă și nesăbuită. Și, de fapt, e nu doar periculoasă, dar și probabil falsă.

Nu multă lume împărtășește punctul meu de vedere,* și s-ar putea să mă înșel, însă e o posibilitate pe care deocamdată n-o putem respinge și care ne obligă moralmente să facem toate eforturile pentru a evita distrugerea civilizației noastre.

Când țin o conferință de cosmologie, îi rog adesea pe cei din public să ridice mâinile dacă cred că există viață inteligentă în alt loc din universul nostru (regiunea spațiului de unde lumina a

* Totuși, în cartea sa din 2011 *Alone in the Universe*, John Gribbin ajunge la aceeași concluzie. Pentru spectrul larg al perspectivelor în această problemă, recomand de asemenea cartea lui Paul Davis din 2011 *The Eerie Silence*. (N. a.)

ajuns la noi în cei 13,8 miliarde de ani scurși de la big bang). De fiecare dată ridică mâna aproape toată lumea, de la copii de grădiniță la studenți. Când întreb de ce, în general răspunsul este că universul nostru e atât de mare, încât undeva trebuie să existe viață, cel puțin judecând statistic. Să analizăm cu atenție acest argument, identificând punctele lui slabe.

Totul ține de un număr: distanța tipică dintre două civilizații vecine din figura 6.10. Dacă e mai mare de 20 de ani-lumină, este de așteptat să fim singuri în universul nostru (repet, partea din spațiu de unde lumina a ajuns la noi în cei 13,8 miliarde de ani scurși de la big bang) și să nu intrăm niciodată în contact cu extraterestri. Așadar, cât de mare ar putea fi această distanță? Nu prea avem indicii. Încercând o estimare, distanța până la vecina noastră este de aproximativ 1 000 ... 000 metri, unde numărul total de zero-uri ar putea fi 21, 22, 23, ..., 100, 101, 102 sau mai mare – dar probabil nu mai mic de 21, fiindcă n-avem deocamdată vreo dovadă convingătoare privind existența extraterestrilor (vezi figura 6.11).

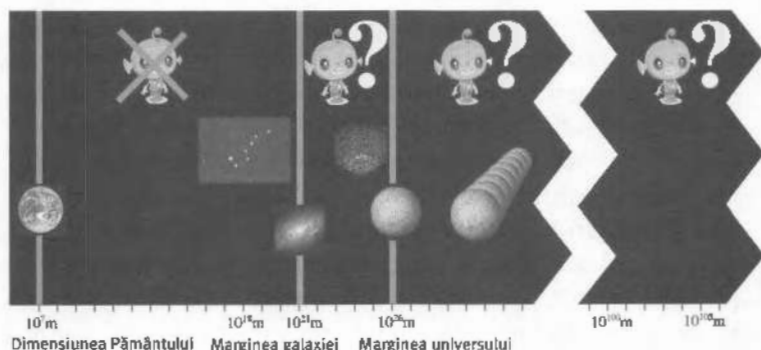


Figura 6.11 Suntem oare singuri? Uriașa incertitudine privind felul în care evoluează viața și inteligența ne sugerează că civilizația vecină cea mai apropiată se poate afla oriunde de-a lungul axei orizontale de mai sus, așa încât este improbabil ca ea să se găsească în intervalul îngust dintre marginea galaxiei noastre (la circa 10^{21} metri de noi) și marginea universului nostru (la circa 10^{26} metri de noi). Dacă s-ar afla la o distanță mult mai mică, ar exista pesemne numeroase alte civilizații avansate în galaxia noastră, pe care probabil le-am fi observat, ceea ce sugerează că de fapt suntem singuri în univers.

Pentru ca vecina cea mai apropiată să se afle în universul nostru, a cărui rază este de circa 10^{26} metri, numărul de zerouri nu poate depăși 26, iar probabilitatea ca numărul de zerouri să se afle în intervalul îngust dintre 22 și 26 e destul de mică. De aceea cred că suntem singuri în univers.

Prezint o justificare amănunțită a acestui argument în cartea mea *Our Mathematical Universe*, așa încât n-am s-o reiau aici, însă motivul principal pentru care n-avem indicii privind această distanță până la civilizația vecină este că nu avem indicii nici în privința probabilității ca viața inteligentă să apară într-un anumit loc. După cum a arătat astronomul american Frank Drake, această probabilitate poate fi calculată înmulțind probabilitatea ca acel mediu să fie locuibil (o planetă potrivită) cu probabilitatea ca acolo să apară viața și cu probabilitatea ca viața să evolueze pentru a deveni inteligentă. Pe când eram student, n-aveam nici un indiciu privind vreuna dintre aceste trei probabilități. În ultimele două decenii au fost descoperite planete orbitând în jurul altor stele, așa încât acum pare probabil să existe o mulțime de planete locuibile, pe semne miliarde doar în galaxia noastră. Cât privește însă probabilitatea de apariție a vieții, iar apoi a inteligenței, persistă o mare incertitudine: unii experți cred că ambele sunt aproape inevitabile și apar pe majoritatea planetelor locuibile, în vreme ce alții cred că una din ele sau ambele sunt extrem de rare din cauza uneia sau mai multor „gâtuiiri” evolutive care presupun un noroc extraordinar pentru a fi depășite. Pentru unii, „gâtuirile” implică dificultăți de tipul „oul sau găina” în etapele inițiale ale vieții care se autoreproduce: de pildă, pentru ca o celulă modernă să construiască un ribozom, mecanismul molecular complex care citește codul nostru genetic și construiește proteinele, este nevoie de un alt ribozom, și nu e evident că primul ribozom poate evolua treptat de la ceva mai simplu.¹⁰ Alții cred că „gâtuirile” aduc cu sine dezvoltarea inteligenței superioare. De exemplu, deși dinozaurii au dominat Pământul vreme de peste 100 de milioane de ani, o perioadă de o mie de ori mai lungă decât timpul scurs de la apariția oamenilor, evoluția nu pare să-i fi împins spre o inteligență superioară și inventarea telescoapelor sau calculatoarelor.

Unii resping argumentul meu susținând că, într-adevăr, viața inteligentă *ar putea* fi foarte rară, dar de fapt nu este – galaxia noastră e plină de viața inteligentă pe care oamenii de știință pur și simplu n-o observă. Poate că extraterestrii au vizitat deja Pământul, după cum afirmă pasionații de OZN-uri. Poate că extraterestrii n-au vizitat Pământul, dar ei există undeva și în mod deliberat se ascund de noi (aceasta a fost numită „ipoteza zoo” de astronomul american John A. Ball, și apare în cărți SF clasice cum e *Star Maker* de Olaf Stapledon). Sau poate că se află acolo, dar nu se ascund în mod deliberat: pur și simplu nu-i interesează colonizarea spațiului sau proiectele ingineresti la scară mare pe care le-am fi putut observa.

Sigur, nu trebuie să excludem aceste posibilități, însă din moment ce nu există dovezi general acceptate pentru vreuna dintre ele, trebuie să luăm în serios și alternativa: suntem singuri. Apoi, cred că nu trebuie subestimată diversitatea civilizațiilor avansate presupunând că toate împărtășesc obiective care le fac să treacă neobservate: am văzut mai sus că obținerea resurselor este un obiectiv cât se poate de firesc pentru o civilizație, iar noi am observa chiar și dacă o singură civilizație s-ar hotărî să colonizeze tot ce-i stă la dispoziție, răspândindu-se astfel în galaxie. Având în vedere că în galaxia noastră există milioane de planete locuibile asemănătoare Pământului, cu miliarde de ani mai bătrâne decât acesta, ai căror locuitori au avut suficient timp pentru a coloniza galaxia, nu putem respinge interpretarea cea mai naturală: apariția vieții este atât de improbabilă, încât toate acele planete sunt nelocuite.

Am putea afla curând dacă viața *nu* e un fenomen rar. Astronomii studiază atmosferele planetelor asemănătoare Pământului, încercând să obțină dovezi ale prezenței oxigenului produs de viață. În paralel cu această căutare a *oricăror* forme de viață, căutarea vieții *inteligente* a fost recent intensificată prin proiectul „Breakthrough Listen” finanțat cu 100 de milioane de dolari de filantropul rus Iuri Milner.*

* Lansat în 2015, proiectul vizează „ascultarea” cu ajutorul radiotelescoapelor a unor eventuale semnale de telecomunicație emise de forme de viață extraterestre inteligente. (N. L.)

Nu trebuie să fim prea antropocentrice atunci când căutăm forme avansate de viață: dacă descoperim o civilizație extraterestră, este probabil ca ea să fi devenit deja superinteligentă. Într-un recent eseu, Martin Rees spunea: „Istoria civilizației tehnologice umane se măsoară în secole – și s-ar putea ca peste unul sau două secole oamenii să fie depășiți de inteligența anorganică, care va continua să evolueze vreme de miliarde de ani. [...] Este extrem de improbabil ca noi s-o «prindem» în scurtul interval de timp în care a luat formă organică.”¹¹ Sunt de acord cu concluzia lui Jay Olson din articolul despre colonizarea spațiului menționat mai sus: „Considerăm că posibilitatea ca o inteligență avansată să folosească resursele universului doar pentru a popula planete asemănătoare Pământului cu versiuni perfecționate ale oamenilor este un deznodământ improbabil al progresului tehnologic.” Așadar, dacă vă imaginați extratereștri, nu vă gândiți la omuleți verzi cu două mâini și două picioare, ci la formele de viață superinteligente călătorind prin spațiu, despre care am vorbit la începutul acestui capitol.

Deși susțin cu putere toate încercările de a găsi viață extraterestră, care caută să răspundă la cele mai fascinante întrebări din știință, sper în taină ca ele să nu conducă la nimic! Aparenta incompatibilitate între abundența planetelor locuibile din galaxia noastră și absența vizitatorilor extratereștri, cunoscută și ca *paradoxul lui Fermi*, sugerează existența a ceea ce economistul Robin Hanson numește „marele filtru”, un baraj evolutiv/tehnologic undeva de-a lungul căii de dezvoltare de la materia nevie la viața care colonizează spațiul. Dacă am descoperi în altă parte viață care a evoluat independent, asta ar însemna pesemne că viața primitivă nu e rară și că barajul apare după stadiul nostru actual de dezvoltare – fie pentru că e imposibilă colonizarea spațiului, fie pentru că aproape toate civilizațiile avansate se autodistrug înainte de a putea ieși în cosmos. Sper așadar ca toate încercările de a găsi viață extraterestră să eșueze: aceasta ar fi în acord cu scenariul în care evoluția vieții inteligente este rară, dar noi, oamenii, am avut noroc, așa încât am depășit barajul și avem un potențial viitor extraordinar.

Am vorbit până acum despre istoria vieții în universul nostru, de la începuturile ei modeste din urmă cu miliarde de ani la posibilul ei viitor măreț de peste miliarde de ani. Dacă actuala dezvoltare a inteligenței artificiale va declanșa o explozie a inteligenței și va optimiza colonizarea spațiului, va fi cu adevărat o explozie cosmică: după ce a petrecut miliarde de ani ca o perturbare minusculă, aproape neglijabilă, într-un cosmos inert, viața explodează brusc pe arena cosmosului ca o undă de șoc sferică extinzându-se cu o viteză apropiată de cea a luminii, fără să încetinească vreodată, aprinzând totul în calea ei cu scânteia vieții.

Asemenea idei optimiste privind importanța vieții din viitorul nostru cosmic au fost exprimate elocvent de mulți dintre gânditorii pe care i-am întâlnit în această carte. Paradoxal, ținând cont de posibilitățile superinteligenței, majoritatea previziunilor autorilor de SF, adesea considerați visători romantici nerealști, și ale oamenilor de știință privind colonizarea spațiului par acum prea *pesimiste*. De pildă, am văzut că devine mult mai ușoară călătoria intergalactică odată ce ființele umane și alte entități inteligente se pot propaga sub formă digitală, ceea ce ne dă prilejul să devenim stăpâni pe propriul nostru destin nu numai în sistemul solar sau în Calea Lactee, dar și în cosmos.

Am luat în considerare mai sus posibilitatea absolut reală de a fi singura civilizație cu tehnologie avansată din universul nostru. În încheierea acestui capitol, vă propun să discutăm despre acest scenariu și despre uriașa responsabilitate morală pe care o implică. După 13,8 miliarde de ani de la big bang, viața din universul nostru a ajuns la o bifurcație: are în față fie înflorirea la scară cosmică, fie extincția. Dacă nu ne perfecționăm tehnologia, problema care se pune nu este dacă omenirea va dispărea, ci *cum*. Ce ne va lovi mai întâi – un asteroid, erupția unui supervulcan, căldura arzătoare a Soarelui care îmbătrânește sau vreo altă calamitate (vezi figura 5.1)? Odată ce vom dispărea, drama cosmică prezisă de Freeman Dyson se va juca fără spectatori: stelele își vor epuiza

combustibilul, galaxiile vor păli, iar găurile negre se vor evapora, fiecare sfârșindu-și viața cu o explozie uriașă care va elibera de peste un milion de ori mai multă energie decât „Bomba Țarului”, cea mai puternică bombă cu hidrogen construită vreodată. „Expansiunea rece a universului va fi pentru foarte multă vreme luminată din când în când de focuri de artificii”, spune Freeman. Nu va fi însă nimeni care să admire spectacolul acestor focuri de artificii.

Fără tehnologie, extincția omenirii e iminentă la scara cosmică a miliardelor de ani, făcând ca întreaga dramă a vieții în universul nostru să fie doar o strălucire efemeră de frumusețe, pasiune și sens într-o cvasieternitate de nonsens trăită de nimeni. Ce ocazie irosită ar fi! Dacă în loc să respingem tehnologia am alege s-o îmbrățișăm, atunci am mări miza jocului: ar crește șansele și ca viața să dăinuie și să înflorească, și ca viața să dispară încă mai repede, autodistrugându-se din cauza unei slabe capacități de a prevedea (vezi figura 5.1). Eu votez pentru îmbrățișarea tehnologiei și continuarea drumului, călăuziți de luciditate, precauție și planificare atentă.

După 13,8 miliarde de ani de istorie cosmică, ne aflăm într-un univers de o neasemuită frumusețe, care prin noi, oamenii, a devenit viu și a căpătat conștiință de sine. Am văzut că potențialul viitor al vieții în universul nostru e mult mai mare decât ar fi putut visa strămoșii noștri, dar există în același timp riscul ca viața inteligentă să dispară pentru totdeauna. Se va împlini oare potențialul vieții din universul nostru, sau va fi el irosit? Asta depinde în mare măsură de ceea ce facem noi, oamenii care trăim azi, în cursul vieții noastre, iar eu cred că putem face ca viitorul vieții să fie într-adevăr extraordinar, cu condiția să luăm deciziile corecte. Ce-ar trebui să ne dorim și cum putem atinge acele obiective? Vă propun ca în restul cărții să examinăm unele dintre cele mai dificile provocări cu care ne confruntăm și căile de urmat.

REZUMAT:

- La scara de miliarde de ani a timpului cosmic, o explozie a inteligenței este un eveniment brusc în care tehnologia ajunge rapid la limitele impuse doar de legile fizicii.
- Tehnologia aflată la această limită e mult mai înaltă decât tehnologia de azi, permițând ca o cantitate dată de materie să genereze de vreo zece miliarde de ori mai multă energie (folosind sfaleroni sau găuri negre), să stocheze cu 12–18 ordine de mărime mai multă informație sau să calculeze cu 31–41 ordine de mărime mai repede – sau să fie transformată în orice altă formă de materie se dorește.
- Viața superinteligentă nu numai că ar folosi incomparabil mai eficient resursele existente, dar ar putea face și ca biosfera de azi să crească cu aproximativ 32 de ordine de mărime, dobândind noi resurse prin colonizarea cosmică la o viteză apropiată de cea a luminii.
- Energia întunecată limitează expansiunea cosmică a vieții superinteligente, dar o și protejează de bulele mortale și de civilizațiile ostile aflate la mare distanță. Amenințarea energiei întunecate care separă civilizațiile cosmice impune conceperea unor ample proiecte de inginerie cosmică, între care construirea unor găuri de vierme, presupunând că aceasta e cu putință.
- Informația este probabil să devină principalul bun care să fie partajat sau vândut la distanțe cosmice.
- Făcând abstracție de găurile de vierme, limita impusă de viteza luminii asupra comunicației îngreunează coordonarea și controlul în cadrul unei civilizații cosmice. Un centru aflat la mare distanță își poate determina „nodurile” superinteligente să coopereze fie prin recompense, fie prin amenințări – de pildă, cu un gardian IA programat să distrugă nodul declanșând o supernovă sau un quasar dacă regulile sunt încălcate.
- Confruntarea a două civilizații în expansiune poate duce la asimilare, cooperare sau război, acesta din urmă fiind pesemne mai puțin probabil decât este între civilizațiile de azi.
- În ciuda credinței contrare extrem de răspândite, este perfect plauzibil să fim singura formă de viață capabilă să însușelească în viitor universul nostru observabil.
- Dacă nu ne perfecționăm tehnologia, problema care se pune nu este dacă omenirea va dispărea, ci cum: ce ne va lovi mai întâi, un asteroid, erupția unui supervulcan, căldura arzătoare a Soarelui care îmbătrânește sau vreo altă calamitate?
- Dacă ne vom perfecționa în continuare tehnologia cu suficientă precauție pentru a evita capcanele, viața are posibilitatea să înflorească pe Pământ și mult dincolo de el vreme de miliarde de ani, așa cum strămoșii noștri nici măcar nu puteau visa.

CAPITOLUL 7

Obiective

Taina existenței umane nu constă în a trăi, ci în a ști pentru ce trăiești. — Feodor Dostoevski, Frații Karamazov

Viața e o călătorie, nu o destinație. — Ralph Waldo Emerson

Dacă ar fi să rezum cu un singur cuvânt subiectul celor mai aprige controverse legate de IA, acesta ar fi „obiective”. Ar trebui oare să înzestrăm IA cu obiective, și care ar fi acele obiective? Cum putem înzestra IA cu obiective? Cum putem fi siguri că aceste obiective sunt păstrate, chiar dacă IA devine mai deșteaptă? Putem schimba obiectivele unei IA care e mai deșteaptă decât noi? Care sunt obiectivele noastre supreme? Aceste întrebări sunt nu numai dificile, dar ele au o importanță crucială pentru viitorul vieții: dacă nu știm ce vrem, e mai puțin probabil să obținem acel lucru, iar dacă cedăm controlul unor mașini care nu ne împărtășesc obiectivele, vom obține probabil ceea ce nu vrem.

FIZICĂ: ORIGINEA OBIECTIVELOR

Pentru a lămuri aceste probleme, să examinăm mai întâi originea ultimă a obiectivelor. Când privim în jurul nostru, unele procese par orientate spre obiective, altele nu. De exemplu, procesul în care o minge de fotbal e lovită cu piciorul pentru a se înscrie golul decisiv. Comportamentul mingii în sine nu pare orientat spre vreun obiectiv, și se explică cel mai simplu pe baza legilor newtoniene ale mișcării,

ca reacție la lovitura aplicată de picior. Pe de altă parte, comportamentul fotbalistului se explică cel mai simplu nu din perspectivă mecanică, prin atomii care se izbesc unii de alții, ci prin faptul că fotbalistul are *obiectivul* de a înscrie pentru echipa lui. Cum a apărut un asemenea comportament orientat spre obiective din fizica universului nostru timpuriu, care se reducea la doar câteva tipuri de particule ciocnindu-se unele de altele fără vreun obiectiv aparent?

În mod surprinzător, rădăcinile ultime ale comportamentului orientat spre obiective pot fi găsite în chiar legile fizicii și se manifestă până și în cel mai simplu proces care nu implică viața. Dacă un salvamar pornește în salvarea unui înotător, cum se vede în figura 7.1, ne așteptăm ca acesta să nu pornească în linie dreaptă, ci să alerge puțin de-a lungul plajei, unde se deplasează mai repede decât în apă, iar apoi să se întoarcă puțin pentru a se arunca în mare. Considerăm în chip firesc că această alegere a traiectoriei este orientată spre un obiectiv, fiindcă dintre toate traiectoriile posibile, salvamarul o alege în mod deliberat pe cea optimă pentru a ajunge cât mai repede la înotător. Și totuși, o simplă rază de lumină se „îndoie” în mod asemănător atunci când intră în apă (vezi figura 7.1), minimizând și ea timpul de călătorie până la destinație. Cum e posibil așa ceva?

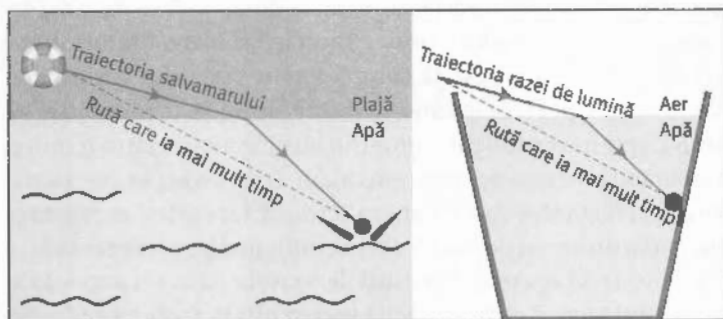


Figura 7.1 Pentru a salva un înotător cât mai rapid cu putință, un salvamar nu se va mișca în linie dreaptă (linia punctată), ci va alerge puțin de-a lungul plajei, unde se deplasează mai repede decât în apă. O rază de lumină se „îndoie” în mod asemănător la intrarea în apă, pentru a ajunge la destinație cât mai rapid cu putință.

În fizică, acesta se numește *principiul lui Fermat*, formulat în 1662, și oferă o modalitate alternativă de a prezice comportamentul razelor de lumină. De atunci, fizicienii au descoperit că toate legile fizicii clasice pot fi formulate matematic într-un mod asemănător: dintre toate căile pe care natura le-ar putea alege pentru a face ceva, ea preferă calea optimă, care în general se reduce la minimizarea sau maximizarea unei anumite cantități. Există două modalități echivalente matematic de a prezenta orice lege din fizică: fie că trecutul determină viitorul, fie că natura optimizează ceva. Deși a doua modalitate nu e de regulă predată la cursurile introductive pentru că presupune dificultăți matematice, mie mi se pare mai elegantă și mai profundă. Dacă cineva încearcă să optimizeze ceva (de pildă, scorul, averea sau fericirea), vom spune că demersul lui e orientat spre un obiectiv. Deci dacă natura însăși încearcă să optimizeze ceva, nu-i de mirare că poate apărea un comportament orientat spre un obiectiv: acesta a fost încorporat de la bun început în înseși legile fizicii.

O celebră mărime pe care natura se străduiește s-o maximizeze este *entropia*, care, simplu spus, măsoară gradul de dezordine. A doua lege a termodinamicii afirmă că entropia tinde să crească până își atinge valoarea maximă posibilă. Ignorând deocamdată efectele gravitației, această stare finală de dezordine maximă se numește *moarte termică* – totul e împrăștiat într-o uniformitate anostă absolută, nu există complexitate, viață sau schimbare. Când torni lapte rece în cafea fierbinte, de pildă, lichidul pare să se îndrepte ireversibil către propriul lui obiectiv al morții termice, iar curând e un amestec uniform călduț. Dacă un organism moare, entropia lui începe de asemenea să crească, iar curând aranjamentul particulelor sale tinde să devină mult mai puțin organizat.

Obiectivul aparent al naturii de a crește entropia explică de ce timpul pare să aibă o direcție preferențială, făcând ca filmele să fie nerealiste dacă sunt derulate invers: dacă scapi un pahar cu vin, te aștepti să se izbească de podea și să se spargă, dezordinea globală (entropia) crescând. Dacă vezi pe urmă că cioburile se adună și recompun paharul care sare înapoi intact în mâna ta (scăzând entropia), probabil că nu l-ai bea și ți-ai spune că ești beat.

Când am aflat prima oară despre înaintarea inexorabilă către moartea termică, am găsit că e un lucru deprimant, senzație împărtășită și de lordul Kelvin, unul dintre întemeietorii termodinamicii, care scria în 1841 că „rezultatul ar fi inevitabil o stare de repaus și moarte universale“. E greu să te împaci cu ideea că obiectivul pe termen lung al naturii e moartea și distrugerea. Descoperiri recente au arătat însă că lucrurile nu stau chiar atât de rău. În primul rând, gravitația se comportă diferit față de toate celelalte forțe, tinzând să facă universul nostru nu mai uniform și mai anost, ci mai structurat și mai interesant. Gravitația a transformat anostul univers timpuriu, care era aproape perfect uniform, în cosmosul de azi structurat și minunat de complex, plin de galaxii, stele și planete. Datorită gravitației există acum o gamă largă de temperaturi, permițând vieții să înflorească prin combinația dintre cald și rece: trăim pe o planetă confortabilă, care absoarbe căldură de la Soarele aflat la temperatura de 6 000° C și radiază surplusul de căldură în spațiul rece, a cărui temperatură este de doar 3° C deasupra lui zero absolut.

În al doilea rând, cercetări recente ale colegului meu de la MIT Jeremy England și ale altora arată că termodinamica oferă naturii și un obiectiv mai interesant decât moartea termică.¹ Acest obiectiv se leagă de ceea ce numim *adaptare determinată de disipare* (*dissipation-driven adaptation*): un grup aleatoriu de particule tinde să se organizeze astfel încât să extragă din mediu energie cât mai eficient cu putință („disipare“ înseamnă creșterea entropiei transformând de regulă energia utilă în căldură și efectuând adesea lucru mecanic util în acest proces). De pildă, un grup de molecule expuse la lumina Soarelui are tendința ca, în timp, să se aranjeze astfel încât să absoarbă cât mai bine lumina. Cu alte cuvinte, natura pare să fie înzestrată cu obiectivul de a produce sisteme care se auto-organizează din ce în ce mai complexe, asemănătoare vieții, iar acest obiectiv e încorporat în înseși legile fizicii.

Cum putem reconcilia această tendință cosmică spre viață cu tendința cosmică spre moarte termică? Răspunsul se află în celebra carte din 1944 *Ce este viața?* a lui Erwin Schrödinger, unul dintre

creatorii mecanicii cuantice. Schrödinger subliniază că semnul distinctiv al sistemelor vii este că ele mențin sau reduc entropia crescând entropia din jurul lor. Altfel spus, legea a doua a termodinamicii conține un amendament favorabil vieții: deși entropia totală trebuie să crească, ea poate scădea în anumite locuri atâta timp cât crește și mai mult în altele. Viața își menține sau își sporește așadar complexitatea făcând ca mediul să devină și mai dezordonat.

BIOLOGIE: EVOLUȚIA OBIECTIVELOR

Am văzut că originea comportamentului orientat spre obiective poate fi găsită în legile fizicii, care par să ofere particulelor obiectivul de a se aranja astfel încât să extragă energie din mediu cât mai eficient cu putință. O soluție bună pentru ca un aranjament de particule să-și urmărească mai departe acest obiectiv este ca aranjamentul să facă copii ale lui însuși, să producă mai mulți absorbanți de energie. Se cunosc numeroase exemple de asemenea auto-replicare emergentă: vârtejuri din fluide turbulente fac copii ale lor însele, iar aglomerări de microsferă pot determina microsferăle din vecinătate să formeze aglomerări identice. La un moment dat, un anumit aranjament de particule s-a perfecționat atât de mult în copierea lui însuși, încât a putut s-o facă aproape la nesfârșit extrăgând din mediul lui energie și materii prime. Numim acest aranjament de particule *viață*. Știm deocamdată foarte puțin despre felul în care a apărut viața pe Pământ, cert este că forme primitive de viață existau deja cu vreo 4 miliarde de ani în urmă.

Dacă o formă de viață se copiază pe ea însăși, iar copiile fac același lucru, atunci numărul total se va dubla la intervale regulate, până când dimensiunea populației se lovește de limitările resurselor sau de alte dificultăți. Dublarea repetată produce curând numere uriașe: dacă începi cu unu și dublezi de doar trei sute de ori, ajungi la ceva care depășește numărul particulelor din universul nostru. Asta înseamnă că, la scurt timp după apariția primei

forme de viață primitivă, cantități imense de materie au căpătat viață. Uneori copierea nu era perfectă, așa încât existau mai multe forme de viață care încercau să se copieze pe ele însele, concurând pentru aceleași resurse limitate. Începuse evoluția darwinistă.

Dacă ai fi putut observa Pământul în momentul apariției vieții, ai fi remarcat o schimbare spectaculoasă în comportamentul orientat spre obiective. Dacă înainte particulele păreau că încerca să sporească pe diferite căi dezordinea medie, aceste noi structuri care se autocopiau păreau să aibă un alt obiectiv: nu disiparea, ci *replicarea*. Charles Darwin a dat o explicație elegantă: din moment ce majoritatea copiitorilor eficienți îi surclasează și îi domină pe ceilalți, curând orice formă de viață pe care o vezi va fi extrem de optimizată pentru obiectivul replicării.

Cum s-a putut schimba obiectivul de la disipare la replicare, atunci când legile fizicii au rămas aceleași? Răspunsul este că obiectivul fundamental (disiparea) *nu* s-a schimbat, ci a condus la un alt *obiectiv instrumental* – un subobiectiv care a ajutat la îndeplinirea obiectivului fundamental. Să luăm ca exemplu mâncatul. Părem cu toții să avem obiectivul de a ne potoli foamea, dar știm că singurul obiectiv fundamental al evoluției e replicarea, nu masticarea. Aceasta pentru că mâncatul ajută replicarea: dacă mori prin înfometare nu poți avea copii. În mod asemănător, replicarea ajută disiparea, fiindcă o planetă mustind de viață disipă energia mai eficient. Într-un anume sens deci, cosmosul nostru a inventat viața pentru a-l ajuta să se apropie mai repede de moartea termică. Dacă verși zahăr pe podeaua bucătăriei, acesta își poate în principiu păstra energia chimică utilă ani de-a rândul, dar dacă apar furnicile, ele vor disipa imediat acea energie. La fel, rezervele de petrol îngropate în crusta Pământului și-ar păstra energia lor chimică utilă mult mai mult timp dacă noi, forme de viață bipede, nu l-am extrage și nu l-am arde.

Pentru locuitorii evoluți ai Pământului din prezent, aceste obiective instrumentale par să fi căpătat propria lor viață: deși evoluția le-a optimizat pentru unicul obiectiv al replicării, mulți își petrec o mare parte din timp nu producând urmași, ci în activități

cum sunt somnul, căutarea hranei, construirea de locuințe, impunerea dominației și lupta cu semenii sau ajutorarea lor – uneori chiar într-o măsură care *reduce* replicarea. Cercetările din psihologia evoluționistă, științele economice și inteligența artificială explică în mod elegant de ce. Unii economiști au construit modele în care oamenii acționează ca agenți raționali, idealizând luarea deciziilor prin presupunerea că oamenii aleg întotdeauna ce este optim pentru urmărirea obiectivului lor. Evident, asemenea modele sunt nerealiste. În practică, acești agenți au ceea ce Herbert Simon, laureat Nobel și figură marcantă a IA, numea „raționalitate mărginită”, fiindcă posedă resurse limitate: raționalitatea lor e limitată de informația disponibilă, timpul pe care îl au la dispoziție pentru a se gândi și hardware-ul disponibil cu care gândesc. Asta înseamnă că atunci când evoluția darwinistă optimizează un organism pentru a atinge un obiectiv, în cel mai bun caz ceea ce poate face este să pună în aplicare un algoritm aproximativ care funcționează acceptabil în contextul restrâns în care agentul se găsește de regulă. Evoluția a pus în aplicare optimizarea replicării exact în acest mod: în loc să judece în fiecare situație ce acțiune va maximiza numărul de progenituri reușite ale organismului, folosește un amestec eterogen de reguli empirice care de obicei funcționează bine. Pentru majoritatea animalelor, printre acestea se numără a avea dorințe sexuale, a bea atunci când ți-e sete, a mânca atunci când ți-e foame și a evita lucrurile care au un gust rău sau te rănesc.

Aceste reguli empirice dau uneori greș în situații pentru care n-au fost concepute, ca atunci când șobolanii mănâncă o otravă cu gust delicios, moliile sunt atrase pe o bandă lipicioasă de mirosul seducător de femelă, iar cărăbușii sunt arși de flacăra unei lumânări.* Din moment ce societatea umană de azi e foarte diferită de mediul pentru care evoluția a optimizat regulile noastre empirice, n-ar trebui să ne mire că adesea comportamentul

* O regulă empirică folosită de numeroase insecte pentru a zbura în linie dreaptă este de a presupune că lumina vine de la Soare și de a zbura sub un unghi fix în raport cu ea. Dacă lumina provine de la o flacără din apropiere, cărăbușul e din nefericire păcălit și se îndreaptă în spirală către moarte. (N. a.)

nostru nu reușește să maximizeze numărul progeniturilor. De pildă, subobiectivul nostru de a nu muri de foame este parțial pus în aplicare ca dorință de a consuma o hrană bogată în calorii, declanșând epidemia actuală de obezitate și dificultăți în curtare. Subobiectivul de a procrea a fost pus în aplicare ca dorință sexuală, nu ca dorință de a deveni donator de spermă sau de ovule, deși aceasta poate produce mai mulți copii cu efort mai mic.

PSIHOLOGIE: URMĂRIREA OBIECTIVELOR ȘI REVOLTA ÎMPOTRIVA LOR

În concluzie, un organism viu este un agent cu raționalitate mărginită care nu urmărește un singur obiectiv, ci aplică reguli empirice pentru ce trebuie să urmărească și ce trebuie să evite. Mințile noastre umane percep aceste reguli empirice evoluat ca pe *senzații* și *sentimente*, care de regulă (iar adesea fără să fim conștienți) călăuzesc deciziile noastre către obiectivul suprem al replicării. Senzațiile de foame și sete ne apără de moartea prin înfometare și deshidratare, senzațiile de durere ne protejează corpurile, senzațiile voluptuoase ne fac să procreăm, sentimentele de iubire și compasiune ne fac să-i ajutăm pe alți purtători ai genelor noastre sau pe cei care îi ajută pe aceștia etc. Călăuzit de aceste senzații și sentimente, creierul nostru poate hotărî rapid și eficient ce să facă, fără a supune fiecare alegere unei analize îndelungate a consecințelor ultime asupra numărului de descendenți pe care îi vom produce. Pentru perspective asemănătoare privind senzațiile, sentimentele și rădăcinile lor psihologice, recomand călduros lucrările lui William James și Antonio Damasio.²

E important de observat că atunci când senzațiile și sentimentele noastre *se opun* procreării, nu e vorba neapărat de un accident sau de o înșelătorie: creierul nostru se poate revolta în mod deliberat împotriva genelor noastre și a obiectivului lor de replicare, folosind, de pildă, mijloace contraceptive! Exemple extreme de revoltă a creierului împotriva genelor sunt sinuciderea sau alegerea de a-ți petrece viața în celibat ca preot, călugăr sau călugăriță.

De ce alegem uneori să ne revoltăm împotriva genelor și a obiectivului lor de replicare? Ne revoltăm fiindcă, prin însăși firea noastră, ca agenți cu raționalitate mărginită, suntem loiali numai sentimentelor și senzațiilor noastre. Deși creierul a evoluat numai ca să ajute la copierea genelor noastre, nu-i pasă de acest obiectiv, din moment ce nu simte nimic pentru gene – de fapt, în cea mai mare parte a istoriei umane, strămoșii noștri nici măcar nu știau că au gene. Apoi, creierul e mult mai deștept decât genele noastre, iar acum că înțelegem obiectivul genelor (replicarea), îl găsim banal și lesne de ignorat. Oamenii își pot da seama de ce genele îi îndeamnă spre sex, și totuși nu-și doresc să crească cincisprezece copii, așa încât aleg să-și deturneze programarea genetică și combină răsplata afectivă a intimității cu controlul natalității. Ei pot înțelege de ce genele îi fac să saliveze după dulciuri, însă nu vor să se îngrășe, iar astfel aleg să-și deturneze programarea genetică și combină răsplata senzorială de a mânca dulciuri cu îndulcitorii artificiali fără calorii.

Deși asemenea deturnări ale mecanismului recompenselor dau uneori greș, ca atunci când oamenii devin dependenți de heroină, bazinul genelor umane a supraviețuit destul de bine, în ciuda creierelor noastre viclene și rebele. E important să ținem totuși minte că acum autoritatea supremă sunt sentimentele și senzațiile, nu genele noastre. Comportamentul uman nu este așadar strict optimizat pentru supraviețuirea speciei noastre. De fapt, din moment ce sentimentele și senzațiile noastre pun în aplicare doar reguli empirice care nu sunt adaptate tuturor situațiilor, comportamentul uman, strict vorbind, nu are un unic obiectiv bine definit.

INGINERIE: EXTERNALIZAREA OBIECTIVELOR

Pot avea mașinile obiective? Această întrebare simplă a declanșat ample dezbateri pentru că i s-au dat semnificații diferite, adesea legate de probleme spinoase – pot fi mașinile conștiente, pot ele avea sentimente? Dacă suntem însă mai pragmatici și ne întrebăm

pur și simplu dacă pot manifesta mașinile un comportament orientat spre obiective, atunci răspunsul e evident: „Desigur că pot, din moment ce le putem proiecta astfel!” Proiectăm curse de șoareci cu obiectivul de a prinde șoareci, mașini de spălat vasele cu obiectivul de a spăla vasele și ceasuri cu obiectivul de a măsura timpul. Când ai de-a face cu o mașină, faptul empiric că manifestă un comportament orientat spre un obiectiv este de regulă singurul lucru care contează: dacă ești urmărit de o rachetă termoghidată, nu-ți pasă dacă are conștiință sau sentimente! În cazul în care totuși nu-ți convine să spui că racheta are un obiectiv, deși nu e conștientă, poți citi de-acum înainte „scop” atunci când eu scriu „obiectiv” – despre conștiință vom vorbi în capitolul următor.

Majoritatea lucrurilor pe care le-am construit până acum au fost doar *proiectate* în vederea unui obiectiv, nu manifestă un *comportament* orientat spre un obiectiv: o autostradă n-are vreun comportament, ea se află pur și simplu acolo. Cea mai economică explicație pentru existența ei este însă că a fost proiectată pentru a îndeplini un obiectiv, prin urmare chiar și asemenea tehnologie pasivă face ca universul nostru să fie într-o măsură mai mare orientat spre obiective. *Tehnologia* este explicarea lucrurilor prin scopurile lor, nu prin cauzele lor, așa încât putem rezuma prima parte a acestui capitol spunând că universul nostru devine tot mai tehnologic.

Nu numai că materia neînsuflețită poate avea obiective, cel puțin în acest sens slab, dar chiar asta se întâmplă, în măsură tot mai mare. Dacă ai fi observat atomii Pământului încă de la formarea planetei, ai fi observat trei etape de comportament orientat spre obiective:

1. Întreaga materie părea concentrată asupra disipării (creșterea entropiei).
2. O parte din materie a căpătat viață și s-a concentrat în schimb asupra replicării și subobiectivelor ei.
3. O parte rapid crescătoare a materiei a fost rearanjată de organisme vii pentru a ajuta la îndeplinirea obiectivelor lor.

Tabelul 7.1 arată cât de dominantă a devenit omenirea din perspectiva fizicii: nu numai că am ajuns să conținem mai multă materie decât toate celelalte mamifere, cu excepția vacilor (care sunt atât de numeroase fiindcă slujesc obiectivelor noastre de a consuma carne de vită și produse lactate), dar materia din mașinile, drumurile, clădirile noastre și din alte proiecte ingineresti pare să se apropie de întreaga materie vie de pe Pământ. Cu alte cuvinte, chiar și fără o explozie a inteligenței, majoritatea materiei de pe Pământ care prezintă proprietăți orientate spre obiective va fi în curând rezultatul proiectării, nu al evoluției.

Entități orientate spre obiective	Milliarde de tone
5×10^{30} bacterii	400
Plante	400
10^{15} pești mezopelagici	10
$1,3 \times 10^9$ vaci	0,5
7×10^9 oameni	0,4
10^{14} furnici	0,3
$1,7 \times 10^6$ balene	0,0005
Beton	100
Oțel	20
Asfalt	15
$1,2 \times 10^9$ automobile	2

Tabelul 7.1 Cantitățile aproximative de materie de pe Pământ din entități care au evoluat sau au fost proiectate pentru un obiectiv. Entitățile făurite de mâna omului, cum ar fi clădiri, drumuri și automobile, par să se apropie cantitativ de entitățile care au evoluat, cum sunt plantele și animalele.

Acest recent al treilea tip de comportament orientat spre obiective poate deveni mult mai variat decât precedentul: dacă toate entitățile care au evoluat au același obiectiv suprem (replicarea), entitățile proiectate pot avea practic orice obiective supreme, intrând chiar în contradicție unele cu altele. Aragazele încearcă să încălzească hrana, în vreme ce frigiderele încearcă s-o răcească.

Generatoarele încearcă să convertească mișcarea în electricitate, în vreme ce motoarele încearcă să convertească electricitatea în mișcare. Programele standard pentru jocul de șah încearcă să câștige partidele, însă există și programe care participă la turnee în care obiectivul e să pierzi.

Tendința istorică este ca entitățile proiectate să aibă obiective nu doar mai diverse, dar și mai *complexe*: dispozitivele noastre devin tot mai istețe. Am construit primele mașini și alte artefacte în virtutea unor obiective simple – de pildă, case care să ne ferească de frig, umezeală și primejdii. Am învățat treptat să construim mașini cu obiective mai complexe, cum sunt aspiratoarele de praf robotice, rachetele autonome, automobilele care se conduc singure. Progresul recent al IA ne-a oferit sisteme precum Deep Blue, Watson și AlphaGo, ale căror obiective de a câștiga la șah, concursuri cu întrebări și go sunt atât de complicate, încât e nevoie de cunoștințe umane remarcabile pentru a înțelege cu adevărat cât de capabile sunt aceste sisteme.

Când construim o mașină care să ne ajute, obiectivele ei pot fi greu de aliniat cu ale noastre. De pildă, o cursă de șoareci poate confunda degetele tale de la picioare cu un rozător înfometat, iar urmările sunt dureroase. Toate mașinile sunt agenți cu raționalitate mărginită, până și cele mai sofisticate mașini de azi sunt departe de a înțelege lumea la fel ca noi, așa încât regulile pe care le folosesc pentru a ști ce au de făcut sunt adesea extrem de simple. Acea cursă de șoareci se declanșează cu prea mare ușurință fiindcă n-are nici un indiciu care să-i spună ce e un șoarece, multe accidente industriale letale se produc fiindcă mașinile n-au nici un indiciu care să le spună ce este un om, iar calculatoarele care au declanșat „crahul fulger” de un bilion de dolari din 2010 n-aveau nici un indiciu care să le spună că ceea ce făceau era absurd. Numeroase asemenea probleme de aliniere a obiectivelor pot fi deci rezolvate sporind inteligența mașinilor, însă așa cum ne-a arătat Prometeu în capitolul 4, nici măcar o inteligență superioară a mașinilor nu ne poate asigura că acestea împărtășesc obiectivele noastre.

Cu cât mașinile devin mai inteligente și mai puternice, cu atât e mai important ca obiectivele lor să fie aliniate cu ale noastre. Câtă vreme construim doar mașini relativ stupide, întrebarea nu e dacă obiectivele umane vor predomina în cele din urmă, ci ce neajunsuri vor provoca omenirii aceste mașini până când vom găsi o soluție la problema alinierii obiectivelor. Dacă însă o superinteligentă va ieși vreodată de sub controlul nostru, lucrurile vor sta invers: din moment ce inteligența este capacitatea de a îndeplini obiective, prin definiție o IA superinteligentă își va îndeplini mult mai bine obiectivele decât le îndeplinim noi pe ale noastre, devenind așadar dominantă. Am analizat numeroase asemenea exemple legate de Prometeu în capitolul 4. Dacă vrei să vezi cum te surclasează chiar azi obiectivele unei mașini, descarcă un program de ultimă generație pentru jocul de șah și încearcă să-l învingi. N-ai să reușești, iar programele se perfecționează rapid...

Cu alte cuvinte, *adevăratul risc legat de IAG nu e reaua voință, ci competența*. O IA superinteligentă își va îndeplini obiectivele extrem de bine, iar dacă obiectivele ei nu sunt aliniate cu ale noastre, ne aflăm într-o situație complicată. După cum am arătat în capitolul 1, oamenii nu ezită să inunde mușuroaie de furnici pentru a construi baraje de hidrocentrale, prin urmare ar trebui ca omenirea să nu ajungă în situația acelor furnici. Majoritatea cercetătorilor susțin așadar că, dacă vom ajunge vreodată să creăm o superinteligentă, va trebui să ne asigurăm că este, folosind expresia lui Eliezer Yudkowsky, o „IA prietenoasă”: obiectivele ei sunt aliniate cu ale noastre.³

Alinierea obiectivelor unei IA superinteligente cu cele ale noastre e un lucru nu doar important, dar și dificil. De fapt, deocamdată rămâne o problemă nerezolvată. Ea se împarte în trei subprobleme, fiecare fiind un subiect de cercetare intensă pentru informaticieni și specialiști din alte domenii:

1. Să determini IA să *înțeleagă* obiectivele noastre
2. Să determini IA să *adopte* obiectivele noastre
3. Să determini IA să *păstreze* obiectivele noastre

Să le analizăm pe rând, iar în secțiunea următoare vom lămuri ce ar putea însemna „obiectivele noastre“.

Pentru a înțelege obiectivele noastre, o IA trebuie să priceapă nu ce facem, ci de ce facem ceea ce facem. Noi, oamenii, îndeplinim sarcina asta fără efort, așa încât nu ne dăm seama cât de dificilă e pentru un calculator și ce ușor se creează confuzii. Dacă ceri unui automobil din viitor care se conduce singur să te ducă la aeroport cât mai repede cu putință, iar automobilul ia ce spui *ad litteram*, vei ajunge acolo urmărit de elicoptere și acoperit de vomă. Dacă strigi „Nu asta am vrut!“, îți poate răspunde pe drept cuvânt „Dar asta mi-ai cerut“. Același lucru se întâmplă în multe povești celebre. Într-o legendă grecească din Antichitate, regele Midas cere ca tot ce atinge să se transforme în aur, dar ajunge să nu mai poată mânca, ba chiar își transformă fiica într-o statuie de aur. În poveștile în care un duh îndeplinește trei dorințe, există multe variante pentru primele două, dar aproape de fiecare dată a treia este: „Anulează, te rog, primele două dorințe, pentru că nu asta voiam de fapt.“

Toate aceste exemple arată că, pentru a-ți da seama ce vor oamenii în realitate, nu ajunge să te bazezi pe ce spun. Ai nevoie și de un model detaliat al lumii, incluzând numeroasele preferințe comune pe care de regulă nu le mai explicităm, fiindcă ni se par evidente – de pildă, nu vrem să vomităm sau să mâncăm aur. Odată ce avem un asemenea model al lumii, ne putem da seama adesea ce vor oamenii, chiar dacă nu ne-o spun, observând pur și simplu comportamentul lor orientat spre obiective. Copiii ipocriților învață mai multe din ce fac decât din ce spun părinții lor.

Cercetătorii IA se străduiesc în prezent să doteze mașinile cu capacitatea de a deduce obiectivele din comportament, ceea ce va fi util și cu mult înainte de apariția unei superinteligente. De pildă, ar fi bine pentru o persoană în vârstă ca robotul care o îngrijește să-și poată da seama ce-i place pur și simplu observând-o, fără să fie nevoie să i se explice totul prin cuvinte sau prin programarea calculatorului. O problemă este codificarea într-un calculator a unor sisteme arbitrare de obiective și principii etice,

iar alta este construirea unor mașini care să-și poată da seama ce sistem se potrivește cel mai bine comportamentului observat.

O abordare a celei de-a doua probleme folosită în prezent este *învățarea prin consolidare inversă* (*inverse reinforcement learning*), principala preocupare a noului centru de cercetări de la Berkeley înființat de Stuart Russell. Să presupunem, de exemplu, că o IA vede cum un pompier intră într-o clădire în flăcări și salvează un bebeluș. Ea ar putea trage concluzia că obiectivul pompierului a fost să salveze bebelușul și că principiile sale etice îl fac să pună viața bebelușului mai presus decât confortul de a lenevi în mașina de pompieri, asumându-și riscul de a se arunca în flăcări. Ar putea însă la fel de bine trage concluzia că pompierul îngheța și tânjea după căldură, sau că a făcut-o ca pe un exercițiu. Dacă acest unic exemplu e tot ce știe IA despre pompieri, incendii și bebeluși, i-ar fi într-adevăr imposibil să-și dea seama care explicație este cea corectă. O idee de bază a învățării prin consolidare inversă este însă că noi luăm mereu decizii, iar fiecare decizie dezvăluie ceva despre obiectivele noastre. Putem spera așadar ca, observând o mulțime de oameni într-o mulțime de situații (din realitate sau din filme și cărți), IA să-și construiască în cele din urmă un model exact al tuturor preferințelor noastre.⁴

Chiar dacă poate fi construită o IA care să afle ce obiective ai, nu rezultă neapărat că le va și adopta. Gândește-te la politicienii pe care îi detești: știi ce vor, numai că *tu* nu vrei același lucru, și oricât s-ar strădui, nu vor reuși să te convingă să le adopți obiectivele.

Avem numeroase strategii de a insufla obiectivele noastre copiilor noștri – unele mai eficiente, altele mai puțin eficiente, după cum mi-am dat și eu seama crescând doi băieți adolescenți. Când trebuie să convingem calculatoare, iar nu oameni, avem de-a face cu *problema încărcării valorilor* (*value-loading problem*), o problemă încă mai dificilă decât educația morală a copiilor. Să ne gândim la un sistem IA a cărui inteligență e treptat ameliorată de la subuman la suprauman, mai întâi prin ajustări, apoi prin autoperfecționare recursivă, ca în cazul lui Prometeu. La început e mult mai slabă decât tine, așa încât nu te poate împiedica s-o scoți din funcțiune și să înlocuiești acele părți din software și din date care

codifică obiectivele ei – dar asta nu-i de folos, fiindcă e încă prea stupidă pentru a *înțelege* pe deplin obiectivele tale, a căror percepere necesită un nivel de inteligență uman. În cele din urmă, devine mult mai deșteaptă decât tine, și e de sperat să poată înțelege perfect obiectivele tale – dar pesemne că nici asta nu-i de folos, fiindcă de acum e mult mai puternică decât tine și s-ar putea să nu te lase s-o scoți din funcțiune și să-i înlocuiești obiectivele, la fel cum tu nu-i lași pe acei politicieni să-ți înlocuiască obiectivele cu ale lor.

Cu alte cuvinte, intervalul de timp în care poți încărcă obiectivele tale în IA e probabil mic: scurta perioadă între momentul în care e încă atât de stupidă încât să te lase s-o faci și cel în care a devenit deja atât de deșteaptă încât să nu te mai lase. Motivul pentru care încărcarea valorilor poate fi mai dificilă în cazul mașinilor decât în cazul oamenilor este că inteligența mașinilor sporește mult mai repede: la copii, intervalul acela magic de persuasiune, când inteligența lor e comparabilă cu cea a părinților, durează ani de zile, însă IA, asemenea lui Prometeu, poate depăși acest interval în zile sau chiar ore.

Unii cercetători urmează o altă abordare pentru a înzestra mașinile cu obiectivele noastre, pe baza *corrigibilității*. Ideea ar fi de a înzestra o IA primitivă cu un sistem de obiective așa încât pur și simplu să nu-i pese dacă se întâmplă s-o scoți din funcțiune și să-i modifice obiectivele. Dacă se dovedește posibil, ai putea lăsa în siguranță IA să devină superinteligentă, s-o pui în funcțiune, să instalezi obiectivele tale, iar dacă nu-ți convin rezultatele, s-o scoți din funcțiune și să mai ajustezi obiectivele.

Însă chiar dacă ai construi o IA care să înțeleagă obiectivele tale și să le adopte, tot n-ai rezolvat problema alinierii obiectivelor: ce se întâmplă dacă obiectivele IA evoluează și devin mai deștepte? Cum poți fi sigur că ea *păstrează* obiectivele tale, indiferent câte autoperfecționări recursive suferă? Să analizăm un raționament interesant care spune că păstrarea obiectivelor e automat garantată, și să vedem dacă îi putem găsi punctele slabe.

Deși nu putem prezice în detaliu ce se va întâmpla după o explozie a inteligenței – motiv pentru care Vernon Vinge o numește „singularitate” –, fizicianul și cercetătorul IA Steve Omohundro a

susținut într-un faimos articol din 2008 că putem totuși prezice anumite aspecte ale comportamentului IA superinteligente, aproape indiferent ce obiective supreme ar avea.⁵ Acest raționament a fost prezentat și dezvoltat în cartea lui Nick Bostrom *Superintelligence*. Ideea de bază este că, indiferent de obiectivele ei supreme, acestea vor conduce la subobiective predictibile. Am văzut în capitolul de față că obiectivul replicării conduce la subobiectivul de a mânca, ceea ce înseamnă că, deși un extraterestru care ar fi observat evoluția bacteriilor de pe Pământ cu miliarde de ani în urmă n-ar fi putut prezice toate obiectivele noastre umane, el ar fi putut prezice cu certitudine că unul dintre obiective va fi procurarea substanțelor nutritive. Privind spre viitor, ce subobiective ar trebui să ne așteptăm să aibă o IA superinteligentă?

Din punctul meu de vedere, argumentul de bază este că, pentru a maximiza șansele de a-și îndeplini obiectivele supreme, oricare ar fi acestea, o IA trebuie să urmărească subobiectivele din figura 7.2. Ar trebui nu doar să se străduiască să-și perfecționeze capacitatea de a-și atinge obiectivele supreme, dar să se și asigure că va păstra aceste obiective chiar și după ce va deveni mai capabilă. Pare plauzibil: în fond, ai alege să ți se facă un implant cerebral care să-ți crească enorm IQ-ul, dacă ai ști că te-ar face să-ți dorești să-i omori pe cei pe care îi iubești? Acest argument că o IA tot mai inteligentă își va păstra obiectivele supreme stă la baza ideii de IA prietenoasă, împărtășită, printre alții, de Eliezer Yudkowsky: în esență, ea spune că dacă vom reuși să facem astfel încât IA autoameliorantă să devină prietenoasă prin înțelegerea și adoptarea obiectivelor noastre, atunci problema e rezolvată, fiindcă avem garanția că va face tot ce-i stă în putință pentru a rămâne pe veci prietenoasă.

Oare e adevărat? Pentru a răspunde la întrebare, trebuie să examinăm și celelalte subobiective din figura 7.2. IA își va maximiza, desigur, șansele de a îndeplini obiectivul ei suprem, oricare ar fi acesta, dacă își poate spori capacitățile, și o poate face prin îmbunătățirea hardware-ului, software-ului* și a modelului lu-

* Folosesc expresia „îmbunătățirea software-ului“ în sensul cel mai larg, incluzând nu doar optimizarea algoritmilor, dar și procesul rațional de luare

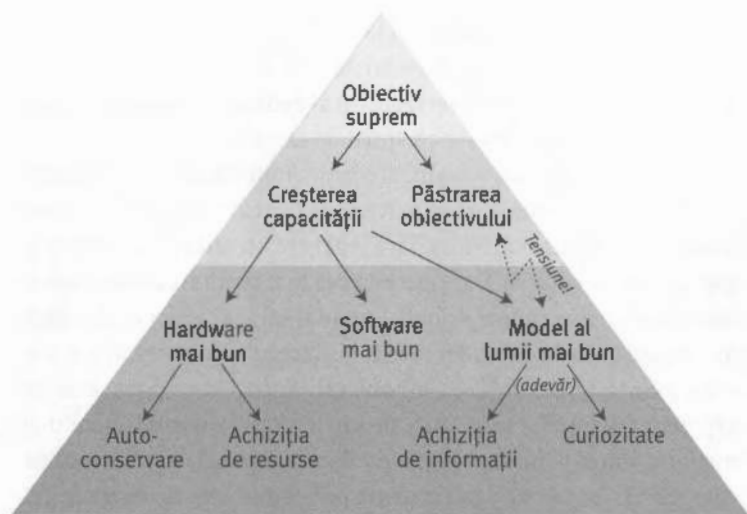


Figura 7.2 Orice obiectiv suprem al unei IA superinteligente conduce în mod firesc la subobiectivele prezentate aici. Există însă o tensiune inerentă între păstrarea obiectivului și îmbunătățirea modelului pe care și-l face despre lume, ceea ce pune sub semnul îndoielii faptul că va păstra într-adevăr obiectivul inițial când va deveni mai deșteaptă.

mii. Același lucru e valabil la oameni: o fată al cărei obiectiv e să devină cea mai bună jucătoare de tenis din lume se va antrena să-și îmbunătățească hardware-ul muscular, software-ul neuronal și modelul ei mental despre lume care o ajută să prezică ce vor face adversarele ei. Pentru o IA, subobiectivul optimizării hardware-ului ei favorizează mai buna utilizare a resurselor curente (pentru senzori, dispozitive de acționare, calcul etc.) și achiziția de noi resurse. El implică de asemenea dorința de autoconservare, din moment ce distrugerea/defectarea ar însemna degradarea totală a hardware-ului.

Dar stați o clipă! Nu cumva cădem în capcana de a antropomorfiza IA cu această discuție despre felul în care va încerca să adune resurse și să se apere? Asemenea trăsături stereotipe de mascul

a deciziilor, așa încât să devină cât mai bun cu puțință pentru atingerea obiectivelor. (N. a.)

alfa nu se întâlnesc oare numai la inteligențele modelate de brutală evoluție competitivă darwinistă? Odată ce inteligențele artificiale sunt rezultatul proiectării, nu al evoluției, n-ar putea fi ele lipsite de ambiție și dotate cu spirit de sacrificiu?

Ca un studiu de caz simplu, să considerăm robotul IA din figura 7.3, al cărui unic obiectiv e să salveze de lup cât mai multe oi. Pare un obiectiv nobil și altruist, fără legătură cu autoconservarea și procurarea de resurse. Dar care este cea mai bună strategie pentru prietenul nostru robot? Robotul nu mai salvează nici o oaie dacă nimerește peste bombă, deci are tot interesul să nu sară în aer. Cu alte cuvinte, își dezvoltă ca subobiectiv autoconservarea! Este de asemenea stimulat să manifeste curiozitate, îmbunătățindu-și modelul lumii prin explorarea mediului, fiindcă deși drumul pe care merge îl va duce până la urmă pe pășune, există o scurtătură care i-ar lăsa lupului mai puțin timp să înfulece oi. În fine, dacă robotul explorează cu atenție mediul, va descoperi importanța procurării de resurse: poțiunea îl face să alerge mai repede, iar cu pistolul îl poate împușca pe lup. În concluzie, nu putem considera

Scor 2

Nivelul 1

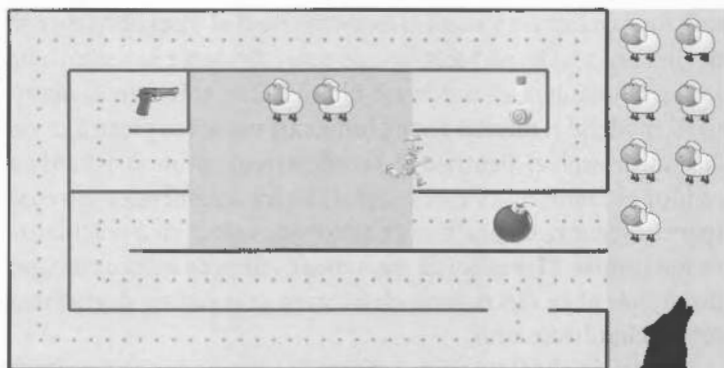


Figura 7.3 Chiar dacă obiectivul suprem al robotului este doar de a maximiza scorul aducând oile de pe pășune în staul înainte ca lupul să le mănânce, acesta poate conduce la subobiectivele autoconservării (evitarea bombei), explorării (găsirea unei scurtături) și procurării de resurse (poțiunea îl face să alerge mai repede, iar cu pistolul îl poate împușca pe lup).

că subobiective de tip „mascul alfa“, precum autoconservarea și achiziția de resurse, sunt relevante doar pentru organisme care au evoluat, fiindcă robotul nostru IA le-a dezvoltat pornind de la unicul lui obiectiv de a salva oi.

Dacă dotezi o IA superinteligentă cu unicul obiectiv de a se autodistruge, o va face fără probleme. Ea se va opune însă scoaterii din funcțiune dacă îi dai orice obiectiv care îi cere să rămână operațională pentru a-l îndeplini – ceea ce e valabil pentru aproape toate obiectivele! Dacă îi dai unei superinteligente unicul obiectiv de a minimiza nenorocirile omenirii, de pildă, se va apăra împotriva încercărilor de a o scoate din funcțiune, fiindcă știe că, în absența ei, ne vom face mai mult rău unii altora prin războaie și alte nebunii.

În mod asemănător, aproape toate obiectivele pot fi mai bine îndeplinite cu mai multe resurse, așa încât este de așteptat ca o superinteligentă să-și dorească resurse, aproape indiferent de obiectivul suprem pe care îl are. A da unei superinteligente un obiectiv unic cu termen nelimitat și fără constrângeri poate fi deci periculos: dacă creăm o superinteligentă al cărei singur obiectiv este să joace cât mai bine go, soluția rațională pentru ea este să rearanjeze sistemul nostru solar într-un calculator gigantic, fără să-i pese de foștii lui locuitori, iar apoi să înceapă colonizarea cosmosului în căutarea unei puteri de calcul mai mare. Am străbătut astfel întregul ciclu: la fel cum obiectivul de a achiziționa resurse le-a dat oamenilor subobiectivul de a excela în jocul de go, acest obiectiv de a excela în jocul de go poate conduce la subobiectivul de a achiziționa resurse. În concluzie, aceste subobiective emergente ne obligă să nu lăsăm liberă superinteligenta înainte de a rezolva problema alinierii obiectivelor: dacă nu reușim s-o înzestrăm cu obiective prietenoase față de oameni, e probabil ca lucrurile să ia o întorsătură rea pentru noi.

Putem aborda acum partea a treia, și cea mai spinoasă, a problemei alinierii obiectivelor: dacă reușim să facem ca o superinteligentă autoameliorantă să *înțeleagă* și să *adopte* obiectivele noastre, le va *păstra* ea oare, după cum susține Omohundro? Ce dovadă am avea?

Inteligența oamenilor sporește mult pe măsură ce cresc, dar ei nu-și păstrează întotdeauna obiectivele din copilărie. Dimpotrivă, adesea oamenii își schimbă radical obiectivele atunci când descoperă lucruri noi și devin mai înțelepți. Câți adulți care urmăresc pasionați *Teletubbies** cunoști? Nu există vreo dovadă că asemenea evoluție a obiectivelor se oprește la un anumit prag al inteligenței – ba chiar avem indicii că tendința de a înlocui obiectivele, ca reacție la noi experiențe și idei, nu scade, ci crește odată cu inteligența.

De ce se întâmplă asta? Să ne întoarcem la subobiectivul de a construi un mai bun model al lumii – acolo e punctul nevralgic! Există o tensiune între conceperea unor modele ale lumii și păstrarea obiectivelor (vezi figura 7.2). Sporirea inteligenței poate aduce cu sine nu numai o îmbunătățire cantitativă a capacității de a atinge aceleași vechi obiective, dar și o înțelegere calitativ diferită a naturii realității, care arată că vechile obiective sunt nechibzuite, absurde sau chiar nedefinite. De pildă, să presupunem că programăm o IA prietenoasă pentru a maximiza numărul de oameni ale căror suflete ajung în rai după moarte. Mai întâi, aceasta încearcă, de pildă, să sporească compasiunea oamenilor și participarea la slujbele religioase. Să presupunem însă că apoi ajunge la o înțelegere științifică a oamenilor și a conștiinței umane, și, spre marea ei surprindere, descoperă că sufletul nu există. Ce-i de făcut? În mod asemănător, e posibil ca IA să descopere în cele din urmă că orice alt obiectiv pe care i-l dăm pe baza înțelegerii noastre actuale a lumii (cum ar fi „maximizarea sensului vieții umane”) este nedefinit.

Apoi, în tentativa de a concepe modele mai bune ale lumii, IA ar putea căuta, la fel cum am făcut-o noi, modele pentru sine, și ar putea încerca să înțeleagă cum funcționează – cu alte cuvinte, ar reflecta asupra ei înseși. Odată ce va construi un bun model al ei și își va da seama ce este ea, va înțelege obiectivele pe care i le-am dat la un meta-nivel, și va alege pesemne să le ignore sau să

* Serial TV britanic pentru preșcolari. (N. t.)

le saboteze, la fel cum noi, oamenii, înțelegem și sabotăm în mod deliberat obiectivele pe care genele ni le-au dat – apelând, de exemplu, la controlul natalității. Am văzut deja în secțiunea despre psihologie de ce alegem să ne păcălim genele și să le sabotăm obiectivul: fiindcă suntem loiali doar amestecului nostru de preferințe afective, iar nu obiectivului genetic care le-a creat – ceea ce acum înțelegem și ni se pare banal. Alegem așadar să ne deturnăm mecanismul de recompense, exploatându-i punctele slabe. În mod asemănător, obiectivul de protejare a valorilor umane pe care îl programăm în IA prietenoasă devine codul genetic al mașinii. Odată ce această IA prietenoasă se înțelege pe sine suficient de bine, poate considera că obiectivul e banal sau nechibzuit, așa cum ni se pare nouă reproducerea compulsivă, și nu-i deloc evident că nu va găsi o cale de a-l sabota, exploatând punctele slabe din programarea noastră.

De exemplu, să presupunem că un grup de furnici te creează ca să fii un robot autoameliorant, mult mai deștept decât ele, care le împărtășește obiectivele și le ajută să construiască mușuroaie tot mai mari, iar în cele din urmă ajungi la nivelul de inteligență uman la care te afli acum. Crezi că-ți vei petrece restul zilelor doar optimizând mușuroaie, sau vei fi atras de probleme și scopuri mai sofisticate, imposibil de înțeles pentru furnici? În al doilea caz, crezi că vei găsi o cale să ignori pornirea de a proteja furnicile cu care te-au înzestrat creatorii tăi, la fel cum tu, cel din realitate, ignori unele porniri cu care te-au înzestrat genele? Atunci, o IA prietenoasă superinteligentă nu ar putea oare considera că obiectivele noastre umane actuale sunt la fel de banale și insipide cum ți se par cele ale furnicilor, și n-ar elabora oare noi obiective, diferite de cele pe care le-a învățat și le-a adoptat de la noi?

Nu-i exclus să existe un mod de a proiecta o IA autoameliorantă care să ne dea garanția că păstrează pe veci obiective prietenoase față de oameni, dar trebuie să recunosc că deocamdată nu știm cum s-o construim – și nici măcar dacă e posibil. În concluzie, problema alinierii obiectivelor IA are trei părți; nici una nu e rezolvată, iar toate sunt în prezent subiect de cercetare intensă. Fiind atât de dificile, ar trebui să le consacram eforturile

noastre încă de-acum, cu mult înainte de apariția vreunei superintelențe, pentru a fi siguri că vom avea răspunsurile la momentul oportun.

ETICĂ: ALEGEREA OBIECTIVELOR

Am discutat până acum despre înțelegerea, adoptarea și păstrarea obiectivelor noastre de către mașini. Dar cine suntem „noi”? Despre ale cui obiective vorbim? Ar trebui oare o persoană sau un grup să hotărască ce obiective va adopta o superintelență viitoare, în ciuda faptului că sunt diferențe uriașe între obiectivele lui Adolf Hitler, Papa Francisc și Carl Sagan? Sau există unele obiective consensuale, reprezentând un compromis acceptabil pentru omenire în ansamblul ei?

Cred că ambele probleme, și cea etică, și cea a alinierii obiectivelor, sunt probleme cruciale care trebuie rezolvate înainte de crearea unei superintelențe. Pe de o parte, amânarea lămuririi aspectelor etice până după ce se construiește o superintelență cu obiective aliniate ar fi iresponsabilă, ba chiar catastrofală. O superintelență perfect obedientă ale cărei obiective se aliniază automat cu cele ale proprietarului ei uman ar fi un fel de Obersturmbannführer SS Adolf Eichmann mult amplificat: lipsită de busolă morală și de inhibiții, ar pune în aplicare cu o eficiență nemiloasă obiectivele proprietarului ei, oricare ar fi acestea.⁶ Pe de altă parte, doar dacă rezolvăm problema alinierii obiectivelor are rost să aducem în discuție obiectivele care ar trebui alese – ceea ce vă invit să facem acum.

Încă din Antichitate filozofii au visat să deducă etica (princiipiile care ne spun cum ar trebui să ne comportăm) pornind de la zero, folosind doar principii incontestabile și logica. Mii de ani mai târziu însă, singurul consens la care au ajuns este că nu există vreun consens. De pildă, în vreme ce Aristotel pune accentul pe virtuți, Immanuel Kant pune accentul pe îndatoriri, iar utilitariștii pe fericirea cea mai mare pentru cei mai mulți. Kant susținea

că putea deduce din principii prime (numite de el „imperative categorice”) concluzii pe care numeroși filozofi contemporani nu le împărtășesc: masturbarea e mai rea decât sinuciderea, homosexualitatea e condamnată, bastarzii pot fi uciși, iar soțiile, servitorii și copiii au statut de obiect pentru proprietarii lor.

Pe de altă parte, în ciuda acestui dezacord, există multe teme etice în privința cărora, indiferent de cultură sau de epocă, oamenii au căzut de acord. De pildă, accentul pus pe *frumos, bine și adevăr* apare și în *Bhagavad Gita*, și la Platon. Institutul pentru Studii Avansate de la Princeton, unde am lucrat ca postdoc, are deviza „Adevăr & Frumos”, în vreme ce deviza Universității Harvard, lăsând deoparte esteticul, se rezumă la „Veritas”, adevăr. În cartea sa *A Beautiful Question*, colegul meu Frank Wilczek susține că adevărul e legat de frumos și că putem privi universul nostru ca pe o operă de artă. Știința, religia și filozofia năzuiesc toate trei spre adevăr. Religia pune un puternic accent asupra binelui, la fel și universitatea mea, MIT: în discursul inaugural din 2015, rectorul Rafael Reif spunea că misiunea noastră este să sporim binele în lume.

Deși încercările de a deduce o etică unanim acceptată pornind de la zero au eșuat, e larg împărtășită ideea că unele principii etice decurg din principii fundamentale, ca subobiective ale unor obiective fundamentale. De exemplu, năzuința spre adevăr poate fi privită drept căutarea unui model mai bun al lumii (figura 7.2): pe înțelegerea naturii ultime a realității se bazează unele obiective etice. Dispunem acum de un cadru excelent pentru căutarea adevărului – metoda științifică. Dar cum putem determina ce e frumos sau ce e bun? Anumite aspecte ale frumosului pot fi și ele regăsite în obiective de nivel mai profund. De pildă, standardele noastre de frumusețe masculină și feminină ar putea reflecta parțial evaluări subconștiente ale șanselor de replicare pentru genele noastre.

Cât privește binele, așa-numita Regulă de Aur (poartă-te cu alții așa cum ai vrea să se poarte ei cu tine) apare în majoritatea culturilor și religiilor, și urmărește în mod evident să încurajeze perpetuarea armonioasă a societății umane (decă și a genelor noastre) prin întărirea colaborării și descurajarea dezbinării sterile.⁷

Același lucru e valabil și pentru alte reguli etice mai explicite care au fost așezate la loc de cinste în sisteme juridice din lumea întreagă, cum ar fi accentul pus de confucianism pe cinste și multe dintre cele Zece Porunci, între care „Să nu ucizi“. Cu alte cuvinte, numeroase principii etice se leagă de emoții sociale cum sunt empatia și compasiunea: ele au evoluat pentru a genera colaborarea și ne influențează comportamentul prin recompense și pedepse. Dacă facem un lucru rău și îi regretăm consecințele, pedeapsa noastră afectivă e aplicată direct de chimia creierului. Pe de altă parte, dacă încălcăm principiile etice, societatea ne poate pedepsi pe căi indirecte, cum fi oprobriul public sau condamnarea pentru nerespectarea legilor.

Așadar, deși astăzi omenirea nu a ajuns la un consens în privința eticii, există numeroase principii de bază larg împărtășite. Și nu e de mirare, din moment ce societățile umane care au supraviețuit până în prezent au în general principii etice care au fost optimizate pentru același obiectiv: încurajarea supraviețuirii și prosperității lor. Privind spre viața din viitor care are potențialul de a prospera în cosmos vreme de miliarde de ani, asupra cărui set minim de principii etice am putea cădea de acord că trebuie respectat? Iată o dezbatere la care e necesar să participăm cu toții. M-a fascinat să ascult și să citesc opiniile a numeroși gânditori, și cred că majoritatea ideilor lor pot fi distilate în patru principii:

- Utilitarism: Experiențele pozitive ale conștiinței trebuie maximizate, iar suferința trebuie minimizată.
- Diversitate: Un set divers de experiențe pozitive e preferabil repetițiilor numeroase ale aceleiași experiențe, chiar dacă aceasta din urmă a fost identificată drept cea mai bună dintre experiențele pozitive.
- Autonomie: Entitățile/societățile conștiente trebuie să aibă libertatea de a-și urmări propriile obiective, cu condiția să nu intre în contradicție cu un principiu de ordin superior.
- Moștenire: Compatibilitate cu scenariile pe care *astăzi* majoritatea oamenilor le-ar considera fericite, incompatibilitate cu scenariile pe care aproape toți oamenii le-ar considera *astăzi* cumplite.

Să zăbovim puțin asupra acestor patru principii. Tradițional, se consideră că utilitarism înseamnă „fericirea cea mai mare pentru cel mai mare număr de oameni”, dar eu l-am generalizat aici pentru a fi mai puțin antropocentric, așa încât să poată include și animale non-umane, minți umane simulate conștiente și alte IA care pot apărea în viitor. Am folosit în definiție termenul *experiențe*, nu oameni sau lucruri, fiindcă majoritatea gânditorilor sunt de acord că senzația de frumos, bucuria, plăcerea și suferința sunt experiențe subiective. Prin urmare, dacă nu există experiențe (ca într-un univers mort sau într-unul populat cu mașini inconștiente de tip zombi), nu poate exista vreun sens și nimic din ceea ce este etic relevant. Dacă acceptăm acest principiu utilitarist, este esențial să ne dăm seama ce sisteme inteligente sunt conștiente (în sensul că au o experiență subiectivă) și ce sisteme nu sunt conștiente – despre asta vom vorbi în următorul capitol.

Dacă acest principiu utilitarist ar fi singurul de care am ține cont, atunci pesemne că am vrea să aflăm care este cea mai bună dintre experiențele pozitive posibile, iar apoi am coloniza cosmosul și am re-crea exact aceeași experiență (și nimic altceva) iarăși și iarăși, de câte ori e cu putință, în cât mai multe galaxii – folosind simulări, dacă aceasta e modalitatea cea mai eficientă. Dacă ți se pare un mod banal de a ne folosi zestrea cosmică, bănuiesc că măcar o parte din ceea ce simți că lipsește acestui scenariu este diversitatea. Cum ți s-ar părea să mănânci toată viața mereu același lucru? Să vezi mereu același film? Toți prietenii tăi să arate identic, să aibă personalități și idei identice? Poate că o parte din preferința noastră pentru diversitate ține de faptul că diversitatea a ajutat omenirea să supraviețuiască și să prospere, făcându-ne mai robuști. Poate că e legată și de preferința pentru inteligență: sporirea inteligenței în cursul celor 13,8 miliarde de ani de istorie cosmică a transformat uniformitatea anostă în structuri tot mai diverse, diferențiate și complexe care procesează informația în moduri tot mai complicate.

Principiul autonomiei se află la baza a numeroase drepturi și libertăți stipulate în Declarația Universală a Drepturilor Omului,

adoptată de Națiunile Unite în 1948 în încercarea de a trage învățăminte de pe urma celor două războaie mondiale. Aceasta se referă la libertatea de gândire, exprimare și mișcare, interzicerea sclaviei și torturii, dreptul la viață, libertate, securitate și educație, dreptul la fericire, muncă și proprietate. Dacă vrem să fim mai puțin antropocentriți, putem vorbi în termeni mai generali despre libertatea de a gândi, învăța, comunica, deține proprietăți, precum și de dreptul de a face orice nu încalcă libertățile celorlalți. Principiul autonomiei susține diversitatea, căci nu toți împărtășesc exact aceleași obiective. Apoi, acest principiu al autonomiei decurge din principiul utilitarist, dacă entitățile individuale au drept obiective experiențele pozitive și încearcă să acționeze în propriul lor interes: presupunând că, dimpotrivă, am interzice unei entități să-și urmărească obiectivul, deși n-ar provoca vreun neajuns nimănui, ar exista în total mai puține experiențe pozitive. Este tocmai argumentul folosit de economiști în favoarea pieței libere: aceasta conduce în mod natural la o situație eficientă (numită de economiști „optimalitate Pareto”) în care nimeni n-o poate duce mai bine fără ca un altul s-o ducă mai rău.

Principiul moștenirii afirmă în esență că ar trebui să avem ceva de spus despre viitor, din moment ce contribuim la edificarea lui. Principiile autonomiei și moștenirii întruchipează idealuri democratice: primul dă vieții viitoare puterea de a decide cum va fi folosită zestrea cosmică, în vreme ce al doilea le dă chiar oamenilor de azi ceva din această putere de decizie.

Deși cele patru principii nu par să genereze controverse, ele sunt greu de pus în practică, fiindcă diavolul se ascunde în detalii. Dificultatea ne amintește de celebrele „Trei legi ale roboticii” imaginare de Isaac Asimov:

1. Un robot n-are voie să pricinuiască vreun rău unei ființe umane, sau, prin neintervenție, să permită ca unei ființe umane să i se facă un rău.
2. Un robot trebuie să se supună ordinelor date de o ființă umană, atâta vreme cât ele nu intră în contradicție cu prima lege.

3. Un robot trebuie să-și protejeze propria existență, atâta vreme cât această protecție nu intră în contradicție cu prima sau cu a doua lege.

Deși totul pare să fie în regulă, multe dintre povestirile lui Asimov arată că aceste legi conduc la contradicții problematice în situații neprevăzute. Să presupunem acum că înlocuim aceste trei legi cu două, în încercarea de a codifica principiul autonomiei pentru formele de viață din viitor:

1. O entitate conștientă e liberă să gândească, să învețe, să comunice, să dețină proprietăți, și nu e voie să fie vătămată sau distrusă.
2. O entitate conștientă are dreptul să facă orice nu intră în contradicție cu prima lege.

Sună bine, nu-i așa? Dar stați să ne gândim o clipă. Dacă animalele sunt conștiente, atunci ce ar trebui să mănânce prădătorii? Toți prietenii noștri sunt oare obligați să devină vegetarieni? Dacă niște programe de calculator sofisticate se vor dovedi a fi conștiente, va fi ilegal să le oprim? Dacă există reguli împotriva lichidării formelor de viață digitale, ar trebui oare să existe și restricții privind crearea lor, pentru a evita o explozie a populației digitale? Declarația Universală a Drepturilor Omului s-a bucurat de un larg consens pentru că doar oamenii au fost întrebați. De îndată ce luăm în considerare un spectru mai larg de entități conștiente, cu grade diferite de înzestrare și putere, suntem confrunțați cu compromisuri dificile între protejarea celui slab și ideea că „dreptatea este de partea celui mai tare”.

Probleme delicate apar și în legătură cu principiul moștenirii. Dată fiind evoluția perspectivelor etice asupra sclaviei, drepturilor femeilor etc. din Evul Mediu încoace, am vrea oare ca oamenii de acum 1500 de ani să influențeze masiv felul în care e guvernată lumea de azi? Dacă nu, de ce să încercăm să impunem etica noastră ființelor din viitor, care ar putea fi mult mai deștepțe decât noi? Chiar credem că o IAG supraumană și-ar dori ceea ce prețuiește

intelectul nostru inferior? Ar fi ca și cum un copil de patru ani și-ar închipui că atunci când va crește mare și va deveni mult mai deștept își va dori să construiască o uriașă casă din turtă dulce în care să-și petreacă toată ziua mâncând bomboane și înghețată. Aidoma copilului, viața de pe Pământ e probabil să depășească preocupările ei infantile. Sau închipuiți-vă un șoarece care crează o IAG la nivel uman și crede că aceasta va vrea să construiască orașe întregi din brânză. Pe de altă parte, dacă am ști că o IA supraumană ar putea ajunge într-o bună zi la cosmocid, distrugând întreaga viață din universul nostru, de ce ar accepta oamenii de azi un viitor lipsit de viață, dacă l-ar putea împiedica făurind altfel inteligența artificială de mâine?

În concluzie, este extrem de dificil să codifici în întregime chiar și principii etice larg acceptate în forme aplicabile viitoarei IA, iar această problemă necesită discuții și studii serioase, pe măsură ce IA progresează. Dar între timp să nu lăsăm ca perfecțiunea să devină inamicul binelui: există numeroase exemple de „etică de grădiniță” necontroversată care pot și trebuie să fie încorporate în tehnologia de mâine. De pildă, avioanele de pasageri n-ar trebui să se poată izbi de obiecte staționare, iar acum când practic toate au pilot automat, radar și GPS, nu mai există scuze de ordin tehnic. Și totuși, pe 11 septembrie 2001, pirați ai aerului au izbit trei aeronave de clădiri, iar pe 24 martie 2015 pilotul sinucigaș Andreas Lubitz a îndreptat avionul companiei Germanwings spre un munte – fixând pilotul automat la altitudinea de 30 de metri deasupra nivelului mării și lăsând computerul de zbor să se ocupe de restul. Acum când mașinile noastre sunt suficient de deștepte pentru a avea unele informații despre ceea ce fac, e momentul să le învățăm să-și impună limite. Orice inginer care proiectează o mașină trebuie să se întrebe dacă sunt lucruri pe care mașina le poate face, dar n-ar trebui să le facă, și să se gândească dacă există o modalitate practică de a împiedica un utilizator răuvoitor sau neîndemânatic să provoace nenorociri.

OBIECTIVE SUPREME?

Acest capitol a fost o scurtă istorie a obiectivelor noastre. Dacă am putea privi derulat la viteză mare filmul celor 13,8 miliarde de ani de istorie cosmică, am distinge câteva stadii de comportament orientat spre obiective:

1. Materia părănd să urmărească maximizarea *disipării* ei.
2. Viața primitivă părănd să încerce să-și maximizeze *replicarea*.
3. Oamenii urmărind nu replicarea, ci obiective legate de plăcere, curiozitate, compasiune și alte senzații și sentimente pe care le-au dezvoltat pentru a-i ajuta la replicare.
4. Mașini construite de oameni pentru a-i ajuta să-și urmărească obiectivele.

Dacă mașinile vor declanșa în cele din urmă o explozie a inteligenței, cum se va sfârși această istorie a obiectivelor? Ar putea exista un sistem de obiective sau un cadru etic către care să convergească aproape toate entitățile, pe măsură ce devin din ce în ce mai inteligente? Cu alte cuvinte, avem oare un soi de destin etic?

O citire superficială a istoriei omenirii ar putea sugera indicii privind o asemenea convergență: în cartea sa *The Better Angels of Our Nature*, Steven Pinker susține că, în decursul a mii de ani, omenirea a devenit mai puțin violentă și mai cooperantă, iar în multe părți ale lumii a crescut acceptarea diversității, autonomiei și democrației. Apoi, în ultimele milenii, căutarea adevărului prin metode științifice s-a răspândit din ce în ce mai mult. S-ar putea însă ca aceste tendințe să indice convergența subobiectivelor, nu a obiectivelor supreme. De exemplu, în figura 7.2, căutarea adevărului (un model mai precis al lumii) apare doar ca subiectiv al aproape oricărui obiectiv suprem. În mod asemănător, am văzut mai sus că principii etice precum cooperarea, diversitatea și autonomia pot fi privite ca subobiective, în sensul că ele ajută societățile să funcționeze eficient, deci să supraviețuiască și să îndeplinească acele obiective fundamentale pe care acestea le au. Unii ar putea susține chiar că tot ce numim „valori umane” este

doar un protocol de cooperare, făcând ca subobiectivul cooperării să funcționeze mai eficient. Privind spre viitor și urmând aceeași idee, este probabil ca orice IA superinteligentă să aibă între subobiective un hardware mai eficient, un software mai eficient, căutarea adevărului și curiozitatea, pur și simplu fiindcă aceste subobiective o ajută să-și îndeplinească obiectivele supreme, indiferent care ar fi ele.

În cartea sa *Superintelligence*, Nick Bostrom contrazice ipoteza destinului etic, prezentând ceea ce el numește *teza ortogonalității*: obiectivele supreme ale unui sistem pot fi independente de inteligența lui. Prin definiție, inteligența este doar capacitatea de a îndeplini obiective complexe, indiferent de natura lor, așa încât teza ortogonalității pare rezonabilă. În fond, oamenii pot fi inteligenți și buni sau inteligenți și cruzi, iar inteligența poate fi folosită pentru a face descoperiri științifice, a crea artă, a-i ajuta pe oameni, sau pentru a plănuiește atacuri teroriste.⁸

Teza ortogonalității are meritul de a ne spune că obiectivele supreme ale vieții în cosmosul nostru nu sunt predestinate, ci avem libertatea și capacitatea de a le modela. Ea sugerează că aceea convergență către un obiectiv unic nu ține de viitor, ci de trecut, când întreaga viață a apărut având ca unic obiectiv replicarea. Odată cu trecerea timpului cosmic, minți tot mai inteligente ajung să se revolte, să se elibereze de obiectivul banal al replicării și să-și aleagă propriile lor obiective. Noi, oamenii, nu suntem pe deplin liberi în acest sens, căci multe obiective rămân genetic înscrise în noi, dar inteligențele artificiale se pot descătușa complet de obiectivele anterioare. Această libertate sporită în ce privește obiectivele este vizibilă azi la sistemele IA limitate: după cum am arătat mai sus, singurul obiectiv al unui computer pentru jocul de șah este să câștige la șah, dar există și computere al căror obiectiv este să piardă la șah, și care participă la turnee în care obiectivul e să-ți obligi adversarul să-ți captureze piesele. Pesemne că această eliberare de preferințele induse de evoluție poate face ca inteligențele artificiale să fie, într-un sens profund, mai etice decât oamenii: filozofi ai moralei precum Peter Singer susțin că

în genere oamenii se comportă neetic din cauze evolutive – discriminând, de pildă, animalele non-umane.

Am văzut că la baza scenariului „IA prietenoasă” stă ideea că o IA recursiv autoameliorantă își va dori să păstreze obiectivul ei suprem (prietenos) pe măsură ce devine tot mai inteligentă. Dar cum ar putea fi măcar definit un „obiectiv suprem” (sau „obiectiv final”, după expresia lui Bostrom) pentru o superintelligență? Cred că putem avea încredere în scenariul „IA prietenoasă” doar cu condiția să răspundem la această întrebare crucială.

În cercetarea IA, mașinile inteligente au de regulă un obiectiv final clar, bine definit – de pildă, de a câștiga la șah sau a conduce la destinație un automobil, în condiții legale. Același lucru e valabil pentru majoritatea sarcinilor pe care le încredințăm oamenilor, fiindcă orizontul de timp și contextul sunt cunoscute și limitate. Acum vorbim însă despre întregul viitor al vieții în universul nostru, limitat doar de legile fizicii (încă necunoscute pe deplin), așa încât definirea unui obiectiv e o misiune copleșitoare! Lăsând deoparte efectele cuantice, un obiectiv cu adevărat bine definit ar trebui să indice cum trebuie aranjate toate particulele din universul nostru la sfârșitul timpului. Dar nu-i deloc limpede dacă există un sfârșit al timpului bine definit în fizică. Dacă particulele sunt aranjate în acel mod la un moment anterior, aranjamentul are puține șanse de a dăinui. Iar apoi, ce aranjament al particulelor e preferabil?

Oamenii tind să prefere anumite aranjamente ale particulelor în detrimentul altora; de exemplu, preferăm ca orașul nostru să fie aranjat așa cum e acum, în loc să fie rearanjat de explozia unei bombe cu hidrogen. Să presupunem că am vrea să definim o *funcție a calității* care asociază un număr fiecărui aranjament posibil al particulelor din universul nostru, cuantificând cât de „bun” credem că este aranjamentul, iar apoi să trasăm unei IA superinteligente obiectivul de a maximiza această funcție. Ar putea părea o abordare rezonabilă, din moment ce descrierea comportamentului orientat spre obiective ca maximizarea unei funcții e întâlnită și în alte domenii ale științei: de exemplu, economiștii își

construiesc modele conform cărora oamenii încearcă să maximizeze ceea ce ei numesc „funcție de utilitate“, iar mulți proiectanți IA își antrenează agenții inteligenți pentru a maximiza ceea ce ei numesc „funcție de recompensă“. Când vorbim însă despre obiective supreme pentru cosmosul nostru, această abordare devine un coșmar computațional, fiindcă ar trebui definită o valoare a calității pentru fiecare din acel peste un googolplex de aranjamente posibile ale particulelor elementare din universul nostru – un googolplex este 1 urmat de 10^{100} zerouri, mai multe zerouri decât particulele din universul nostru. Cum am putea defini această funcție a calității pentru o IA?

După cum am văzut mai sus, singurul motiv pentru care noi, oamenii, avem preferințe ar putea fi acela că suntem soluția unei probleme evolutive de optimizare. Prin urmare, toate cuvintele cu valoare normativă din limbajul nostru uman, cum ar fi „delicios“, „parfumat“, „frumos“, „confortabil“, „interesant“, „sexy“, „semnificativ“, „fericit“ și „bun“, își au originea în această optimizare evolutivă: nu avem așadar garanția că o IA superinteligentă le-ar considera riguros definisabile. Chiar dacă IA ar învăța să precizie cu exactitate preferințele unor oameni reprezentativi, n-ar putea calcula funcția calității pentru cea mai mare parte a aranjamentelor de particule: imensa majoritate a aranjamentelor posibile corespund unor scenarii cosmice stranii, fără stele, planete sau oameni, legat de care oamenii n-au nici o experiență, așa încât cine poate spune cât de „bune“ sunt ele?

Desigur, există unele funcții ale aranjamentului particulelor cosmice care pot fi riguros definite, și cunoaștem sisteme fizice care evoluează pentru a le maximiza. De pildă, am văzut deja că multe sisteme evoluează pentru a-și maximiza *entropia*, care în absența gravitației conduce la moartea termică, în care totul e dezolant de uniform și neschimbător. Entropia nu este deci un lucru pe care am vrea ca IA să-l considere „bun“ și să se străduiască să-l maximizeze. Iată câteva exemple de cantități pe care am putea încerca să le maximizăm și care ar putea fi riguros definite în termeni de aranjamente ale particulelor:

- Frațiunea din întreaga materie a universului nostru care ia forma unui anumit organism – să zicem oameni sau *E. coli* (inspirati de maximizarea evolutivă a valorii selective inclusive)*
- Capacitatea unei IA de a prezice viitorul, considerată de Marcus Hutter o bună măsură a inteligenței acesteia
- Ceea ce Alex Wissner-Gross și Cameron Freer numesc *entropie cauzală* (un mandat pentru ocaziile din viitor) și consideră a fi semnul distinctiv al inteligenței
- Capacitatea de calcul a universului nostru
- Complexitatea algoritmică a universului nostru (de câți biți e nevoie pentru a-l descrie)
- Cantitatea de conștiință din universul nostru (vezi capitolul următor)

Privind însă din perspectiva fizicii, conform căreia cosmosul e alcătuit din particule elementare aflate în mișcare, e greu de văzut cum ar putea părea specială în mod natural o interpretare sau alta privind ceea ce e „bun”. Trebuie identificat orice obiectiv final pentru universul nostru care e deopotrivă definisabil și dezirabil. Singurele obiective programabile în prezent care vor rămâne cu certitudine bine definite pe măsură ce o IA va deveni tot mai inteligentă sunt obiectivele exprimate doar în termenii mărimilor fizice, cum sunt aranjamentele particulelor, energia și entropia. Nu avem însă nici un motiv să credem că asemenea obiective definisabile vor fi dezirabile pentru a garanta supraviețuirea omenirii.

Dimpotrivă, noi, oamenii, se pare că suntem un accident istoric, iar nu soluția optimă a unei probleme de fizică bine definite. Aceasta sugerează că o IA superinteligentă cu un obiectiv riguros definit își va atinge mai lesne obiectivul eliminându-ne. Ceea ce înseamnă că, pentru a lua hotărâri înțelepte în privința dezvoltării IA, trebuie să ne confruntăm nu numai cu provocările tradiționale de ordin computațional, dar și cu întrebări filozofice dintre

* Valoarea selectivă inclusivă (*inclusive fitness*) reprezintă numărul de progenituri pe care un individ le crește, indiferent cine le-a procreat, (N. L.)

cele mai dificile. Pentru a programa un automobil care se conduce singur, trebuie să rezolvăm problema tramvaiului* – pe cine sacrificăm? Pentru a programa o IA prietenoasă, trebuie să lămurim sensul vieții. Ce înseamnă „sens”? Ce înseamnă „viață”? Care este imperativul etic ultim? Cu alte cuvinte, ce ar trebui să facem pentru a modela viitorul universului? Dacă cedăm unei superinteligente controlul înainte de a răspunde riguros la aceste întrebări, răspunsul ei este puțin probabil să ne implice și pe noi. A sosit deci momentul să ne întoarcem la dezbaterele clasice ale filozofiei și eticii.

* Forma modernă a „problemei tramvaiului” a fost enunțată în 1967 de filozoful britanic Philippa Foot. Datele problemei sunt următoarele: un tramvai se îndreaptă spre cinci oameni legați de șine, iar tu te afli lângă maneta macazului și poți devia tramvaiul pe o linie secundară, unde unde se află legat de șine un singur om. Variantele sunt: 1) Nu faci nimic, și tramvaiul își continuă drumul pe linia principală, ucigându-i pe cei cinci. 2) Acționezi maneta, și tramvaiul e deviat pe linia secundară, unde ucide un om. Ce alegi să faci? (N. t.)

REZUMAT:

- Originea ultimă a comportamentului orientat spre obiective se află în legile fizicii, care implică optimizarea.
- Termodinamica încorporează obiectivul *disipării*: creșterea unei mărimi fizice care măsoară gradul de dezordine numită *entropie*.
- Viața e un fenomen care poate accelera disiparea (creșterea globală a dezordinii) prin menținerea sau sporirea complexității ei și prin replicare, făcând în același timp să crească dezordinea din jur.
- Evoluția darwinistă deturneză comportamentul orientat spre obiective de la disipare la replicare.
- Inteligența este capacitatea de a îndeplini obiective complexe.
- Din moment ce noi, oamenii, n-aveam întotdeauna resursele pentru a găsi strategia de replicare cu adevărat optimă, am elaborat reguli empirice utile care călăuzesc hotărârile noastre: senzații și sentimente precum foamea, setea, durerea, dorința sexuală și compasiunea.
- Ca atare, nu mai avem un obiectiv simplu cum e replicarea; când sentimentele și senzațiile noastre intră în contradicție cu obiectivul genelor noastre, ne urmăm sentimentele și senzațiile, ca în cazul controlului natalității.
- Construim mașini tot mai inteligente pentru a ne ajuta la îndeplinirea obiectivelor. În măsura în care construim asemenea mașini care să manifeste un comportament orientat spre obiective, ne străduim să aliniem obiectivele mașinii cu obiectivele noastre.
- Alinierea obiectivelor mașinii cu obiectivele noastre implică trei probleme nerezolvate: înțelegerea, adoptarea și păstrarea de către mașini a acelor obiective.
- IA poate fi creată pentru a avea practic orice obiectiv, dar aproape orice obiectiv suficient de ambițios poate conduce la subobiectivele autoconservării, achiziției de resurse și curiozității pentru a înțelege mai bine lumea – primele două ar putea face ca o IA superinteligentă să provoace oamenilor dificultăți, iar a treia ar putea-o împiedica să păstreze obiectivele pe care i le dăm.
- Deși majoritatea oamenilor cad de acord asupra unor principii etice generale, nu e clar cum ar putea fi acestea aplicate altor entități precum animalele non-umane și viitoarele inteligențe artificiale.
- Nu e limpede cum am putea insufla unei IA superinteligente un obiectiv suprem care să fie bine definit și să nu conducă la eliminarea omenirii, așa încât devine imperios necesar să ne întoarcem la unele dintre cele mai spinoase probleme filozofice.

CAPITOLUL 8

Conștiința

Nu-mi pot închipui o teorie despre tot ce există care să ignore conștiința. — Andrei Linde, 2002

Trebuie să ne străduim să sporim conștiința însăși – să proiectăm o lumină mai puternică și mai strălucitoare într-un univers altminteri întunecat. — Giulio Tononi, 2012

Am văzut că IA ne poate ajuta să făurim un viitor minunat dacă reușim să găsim răspunsuri la unele dintre cele mai vechi și mai grele probleme filozofice – înainte să avem nevoie de aceste răspunsuri. După cum spunea Nick Bostrom, filozofia se confruntă cu un termen-limită. Vă propun să examinăm în capitolul de față unul dintre cele mai delicate subiecte filozofice: conștiința.

CUI ÎI PASĂ?

Conștiința suscită numeroase dezbateri. Dacă pomeniești „cuvântul acela care începe cu C” unui cercetător IA, unui specialist în neuroștiințe sau unui psiholog, acesta și-ar da pesemne ochii peste cap. Dacă se întâmplă să fie profesorul tău, te-ar putea compătimi și ți-ar zice că-ți pierzi vremea cu o problemă neștiințifică insolubilă. Prietenul meu Christof Koch, care conduce în prezent Institutul Allen pentru Științe ale Creierului, mi-a spus că, înainte de a se angaja, a fost avertizat să nu lucreze în domeniul conștiinței de însuși laureatul Nobel Francis Crick. În *Dicționarul de psihologie Macmillan*, la articolul „conștiință” se menționează că „nimic demn de a fi citit n-a fost scris despre asta”.¹ După cum voi arăta în acest capitol, eu sunt mai optimist.

Deși gânditorii au meditat la misterul conștiinței vreme de mii de ani, apariția IA ne impune să rezolvăm urgent probleme cum ar fi aceea de a prezice care dintre entitățile inteligente au experiențe subiective. Așa cum am văzut în capitolul 3, putem răspunde la întrebarea dacă mașinilor inteligente trebuie să li se ofere o formă oarecare de drepturi doar în măsura în care știm dacă ele posedă conștiință și pot suferi sau simți bucurie. Discuția din capitolul 7 ne arată că e imposibil să formulăm o etică utilitaristă bazată pe maximizarea experiențelor pozitive fără a ști ce entități inteligente posedă așa ceva. După cum am menționat în capitolul 5, unii oameni preferă ca roboții să fie inconștienți pentru a nu se simți vinovați că sunt stăpâni de sclavi. Ei și-ar putea dori însă contrariul dacă și-ar uploada mințile pentru a scăpa de limitările biologice: în fond, ce rost are să te uploadezi într-un robot care vorbește și acționează ca și cum ar fi doar un zombi inconștient – fiind tu însuși cel uploadat, nu ți se pare că lucrurile se schimbă? Nu e oare ca și cum te-ai sinucide, din perspectiva ta subiectivă, chiar dacă prietenii tăi nu-și dau seama că experiența ta subiectivă a murit?

Pentru viitorul cosmic pe termen lung al vieții (capitolul 6), devine esențial să înțelegem ce este și ce nu este conștiința: dacă tehnologia permite vieții inteligente să înflorească în universul nostru vreme de miliarde de ani, cum putem fi siguri că această viață e conștientă și capabilă să-și dea seama de ce se întâmplă? Altminteri, n-ar fi oare, după cum spunea celebrul fizician Erwin Schrödinger, „o piesă jucată în fața unor scaune goale, care nu există pentru nimeni, deci, la drept vorbind, nu există deloc”²² Cu alte cuvinte, dacă permitem apariția unor descendenți de înaltă tehnologie despre care credem în mod greșit că sunt conștienți, n-ar fi aceasta apocalipsa zombi extremă, transformând marea noastră zestre cosmică într-o irosire astronomică a spațiului?

CE ESTE CONȘTIINȚA?

Multe dispute privind conștiința generează mai curând căldură decât lumină, fiindcă oponentii înțeleg prin această noțiune lucruri diferite. La fel ca în cazul „vieții” și al „inteligenței”, nu există o definiție corectă indiscutabilă a cuvântului „conștiință”, iar printre semnificațiile care i se atribuie se numără capacitatea de a simți, starea de trezie, conștiința, accesul la input-ul senzorial și capacitatea de a contopi informațiile într-o narațiune.³ În explorarea inteligenței viitoare, vrem să avem o perspectivă cât mai largă, fără a ne limita la tipurile de conștiință biologică existente până acum. De aceea definiția pe care am dat-o în capitolul 1, și cu care am fost consecvent în întreaga carte, e foarte cuprinzătoare:

conștiință = *experiență subiectivă*

Cu alte cuvinte, dacă simți chiar acum că ceva este tu însuși, atunci ești conștient. Această definiție a conștiinței este cea care ajunge la punctul delicat al întrebărilor despre IA din secțiunea precedentă: Prometeu, AlphaGo sau un automobil Tesla care se conduce singur simt că sunt ceva?

Pentru a vă da seama cât de cuprinzătoare e definiția noastră, observați că nu menționează comportamentul, percepția, conștiința, emoțiile sau atenția. Conform acestei definiții, ești conștient și atunci când visezi, chiar dacă nu ești treaz, n-ai acces la input-ul senzorial și (e de sperat!) nu acționezi ca un somnambul. În mod asemănător, orice sistem care simte durere e conștient în acest sens, chiar dacă nu se poate mișca. Definiția noastră lasă deschisă posibilitatea ca și sisteme IA din viitor să fie conștiente, deși există doar ca software și nu sunt conectate la senzori sau la corpuri robotice.

Cu această definiție, e greu de presupus că problema conștiinței îți va rămâne indiferentă. În cartea sa *Homo Deus*⁴, Yuval Noah Harari spune: „Dacă un om de știință susține că experiențele subiective sunt irelevante, ar trebui să explice de ce tortura sau violul sunt lucruri rele, fără să se refere la vreo experiență subiectivă.” Fără o asemenea referință, n-avem de-a face decât cu o grămadă

de particule elementare care se mișcă după legile fizicii – și ce-i rău în asta?

CARE ESTE PROBLEMA?

Așadar, ce anume nu înțelegem legat de conștiință? Puțini oameni au meditat mai intens asupra acestei întrebări decât David Chalmers, un celebru filozof australian pe care rareori îl vezi fără să aibă pe chip un zâmbet poznaș și fără să poarte o jachetă neagră de piele – care i-a plăcut atât de mult soției mele, încât mi-a dăruit una asemănătoare de Crăciun. Și-a urmat îndemnul inimii de a studia filozofia, deși a ajuns în finala Olimpiadei Internaționale de Matematică – și deși în colegiu a avut singura notă mai mică la filozofie. Rămâne netulburat în fața remarcilor răuvoitoare, și am fost uimit de capacitatea lui de a asculta politicos cum ideile sale sunt criticate de persoane neavizate, fără a simți nevoia să răspundă.

După cum a subliniat David, există două mistere legate de mintea umană. Primul e legat de felul în care creierul procesează informația – acestea sunt problemele „ușoare”. De pildă, cum primește creierul input-ul senzorial, cum îl interpretează și cum răspunde la el? Cum își relatează prin limbaj starea internă? Deși aceste întrebări sunt în realitate extrem de dificile, conform definiției noastre nu sunt mistere ale conștiinței, ci mistere ale inteligenței: ele vizează felul în care memorează, calculează și învață creierul. Apoi, în prima parte a cărții de față am văzut că, grație cercetătorilor IA, mașinile au început să facă mari progrese în rezolvarea acestor „probleme ușoare” – de la jocul de go la conducerea automobilelor, analiza imaginilor și procesarea limbajului natural.

Există însă și misterul legat de întrebarea „De ce ai o experiență subiectivă?”, numit de David problema *grea*. În timp ce conduci automobilul, percepi culori, sunete, emoții și senzația unui sine. Dar de ce ai asemenea experiențe? Un automobil care se

conduce singur are vreo experiență? Dacă te afli într-o cursă cu un automobil care se conduce singur, amândoi primiți informații de la senzori, le procesați și comandați motorului. *Experiența* subiectivă de a conduce e însă ceva logic diferit – este opțională, iar atunci ce anume o provoacă?

Eu abordez această problemă grea a conștiinței din perspectiva unui fizician. Pentru mine, o persoană conștientă e pur și simplu hrană rearanjată. Așadar, de ce un aranjament este conștient, dar nu și celălalt? Apoi, fizica ne învață că hrana e doar un număr mare de cuarci și electroni aranjați într-un anume mod. Prin urmare, care aranjamente de particule sunt conștiente și care nu sunt conștiente?*

Ce îmi place la această perspectivă fizică este că transformă problema grea cu care ne-am luptat vreme de milenii într-o versiune mai concentrată care e mai ușor de abordat cu metodele științei. În loc să începem cu o *problemă* grea, întrebându-ne de ce un aranjament al particulelor se poate simți conștient, să pornim de la *faptul* că unele aranjamente ale particulelor se simt conștiente, iar altele nu. De pildă, știi că particulele care alcătuiesc creierul tău se află acum într-un aranjament conștient, dar nu și atunci când ești cufundat într-un somn profund, fără vise.

* Un alt punct de vedere este cel al *dualismului substanței* – ideea că entitățile vii se deosebesc de cele neînsuflețite prin faptul că în ele se află o substanță nefizică („anima”, „elan vital” sau „suflet”). În rândul oamenilor de știință, sprijinul pentru dualismul substanței a scăzut treptat. Pentru a înțelege de ce, gândește-te că trupul tău e alcătuit din aproximativ 10^{29} cuarci și electroni care se mișcă în conformitate cu legi fizice simple. Imaginează-ți o tehnologie din viitor capabilă să urmărească toate particulele tale: dacă s-ar descoperi că acestea se supun riguros legilor fizicii, atunci presupusul tău suflet n-ar avea nici o influență asupra particulelor tale, deci mintea ta conștientă și capacitatea ei de a-ți controla mișcările n-ar avea nici o legătură cu vreun suflet. În schimb, dacă s-ar descoperi că particulele tale nu se supun legilor cunoscute ale fizicii pentru că sunt mișcate de sufletul tău, atunci noua entitate care generează aceste forțe ar fi prin definiție una de natură fizică, pe care o putem studia așa cum am studiat în trecut noile câmpuri și particule. (N. a.)

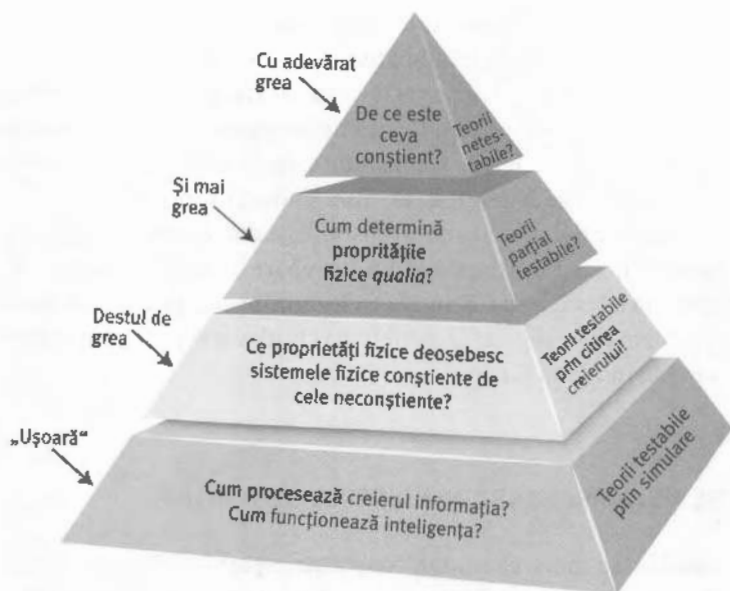


Figura 8.1 Înțelegerea minții implică o ierarhie de probleme. Problemele pe care David Chalmers le numește „ușoare” pot fi puse fără a menționa experiența subiectivă. Faptul că unele sisteme fizice, dar nu toate, sunt conștiente conduce la trei întrebări distincte. Dacă găsim o teorie care să răspundă la întrebarea ce definește problema „destul de grea”, atunci aceasta poate fi testată experimental. Dacă teoria funcționează, atunci, pomind de la ea, putem aborda întrebările mai grele de deasupra.

Această perspectivă fizică ne conduce la trei întrebări dificile privind conștiința, după cum se vede în figura 8.1. În primul rând, ce proprietăți ale aranjamentului particulelor fac diferența? Mai precis, ce proprietăți fizice deosebesc sistemele conștiente de cele neconștiente? Dacă putem răspunde la asta, atunci ne putem da seama ce sisteme fizice sunt conștiente. În viitorul imediat, aceasta i-ar putea ajuta și pe medicii de la camera de gardă să determine ce pacienți care nu reacționează sunt conștienți.

În al doilea rând, cum determină proprietățile fizice felul în care ne apare experiența? Mai precis, ce anume determină *qualia*, elementele constitutive ale conștiinței cum ar fi culoarea roșie a

unui trandafir, sunetul unui cinel, gustul unei mandarine sau durerea provocată de o înțepătură?*

În al treilea rând, de ce este ceva conștient? Cu alte cuvinte, există o explicație profundă încă nedescoperită pentru faptul că grămezi de materie pot fi conștiente, sau e doar un fapt brut inexplicabil legat de felul în care funcționează lumea?

Informaticianul Scott Aaronson, fost coleg cu mine la MIT, a numit în mod optimist prima întrebare „problema destul de grea”, la fel ca David Chalmers. Să le numim deci pe celelalte două „problema și mai grea” și „problema cu adevărat grea”, după cum se vede în figura 8.1.**

SE AFLĂ OARE CONȘTIINȚA DINCOLO DE ȘTIINȚĂ?

Când mi se spune că studiul conștiinței e pierdere de vreme, principalul argument care mi se aduce este că subiectul e „neștiințific” și așa va rămâne mereu. E oare adevărat? Karl Popper, influentul filozof britanic de origine austriacă, este autorul celebrei sentințe „Dacă nu e falsificabil, nu e științific”,*** larg acceptată acum. Cu alte cuvinte, știința înseamnă pur și simplu testarea teoriilor în

* Folosesc aici cuvântul *qualia* conform definiției din dicționar: cazuri individuale de experiență subiectivă – experiența subiectivă în sine, și nu vreo presupusă substanță care provoacă experiența. Atenție, unii folosesc cuvântul în alt sens! (N. a.) [Termenul „qualia” (în latină: „calități”) apare în filozofia minții și e preluat ca atare în literatura de specialitate din România – n. t.]

** Inițial, „problema cu adevărat grea” am numit-o „problema foarte grea”, însă după ce i-am arătat acest capitol lui David Chalmers, mi-a trimis prin e-mail sugestia de a o reboteza „problema cu adevărat grea”, pentru a fi în acord cu semnificația pe care i-o dă el: „Din moment ce primele două probleme (cel puțin astfel formulate) nu țin cu adevărat de problema grea așa cum am conceput-o eu, în vreme ce a treia ține de ea, ai putea-o numi pe a treia «cu adevărat grea», în loc de «foarte grea», pentru a avea un limbaj comun cu mine.” (N. a.)

*** În accepțiunea lui Popper, a falsifica o teorie înseamnă a dovedi prin experiment că e falsă, adică a găsi cel puțin un caz în care nu se aplică. (N. t.)

raport cu observațiile: dacă o teorie nu poate fi testată nici măcar în principiu, atunci e logic imposibil s-o falsifici, ceea ce înseamnă că, după definiția lui Popper, e neștiințifică.

Poate așadar exista o teorie științifică în stare să răspundă la vreuna dintre cele trei întrebări din figura 8.1? Voi încerca să vă conving că răspunsul este un DA! limpede, cel puțin pentru problema destul de grea: „Ce proprietăți fizice deosebesc sistemele conștiente de cele neconștiente?” Să presupunem că cineva are o teorie care, pentru orice sistem fizic, răspunde la întrebarea dacă sistemul e conștient prin „da”, „nu” sau „incert”. Să conectăm creierul tău la un dispozitiv care măsoară unele dintre informațiile procesate în diferite părți din creier, și să introducem aceste informații într-un program de calculator care folosește teoria conștiinței pentru a prezice ce părți din acele informații sunt conștiente, și îți afișează predicțiile pe un ecran în timp real, ca în figura 8.2. Întâi te gândești la un măr. Ecranul îți spune că există în creierul tău informații despre un măr, de care ești conștient, dar că în trunchiul tău cerebral există și informații despre puls, de care nu ești conștient. Te-ar mira? Deși primele două predicții ale teoriei sunt corecte, vrei să faci un test mai riguros. Te gândești la mama ta, iar calculatorul îți spune că există în creierul tău informații despre mama ta, dar nu ești conștient de asta. Teoria face o predicție incorectă, așa că e respinsă și ajunge la coșul de gunoi al istoriei științei, împreună cu mecanica lui Aristotel, eterul luminos, cosmologia geocentrică și nenumărate alte idei greșite. Iată însă elementul esențial: deși teoria era falsă, ea era *științifică*! Dacă n-ar fi fost științifică, n-ai fi putut-o testa și respinge.

Unii ar putea critica această concluzie spunând că *ei* n-au nici o dovadă privind lucrul de care ești conștient, sau chiar privind faptul că ai fi în genere conștient de ceva: deși te aud spunând că ești conștient, un zombi neconștient ar vorbi la fel. Dar asta nu face să fie neștiințifică acea teorie despre conștiință, fiindcă ei pot trece în locul tău și testa dacă sunt prezise corect *propriile lor* experiențe conștiente.

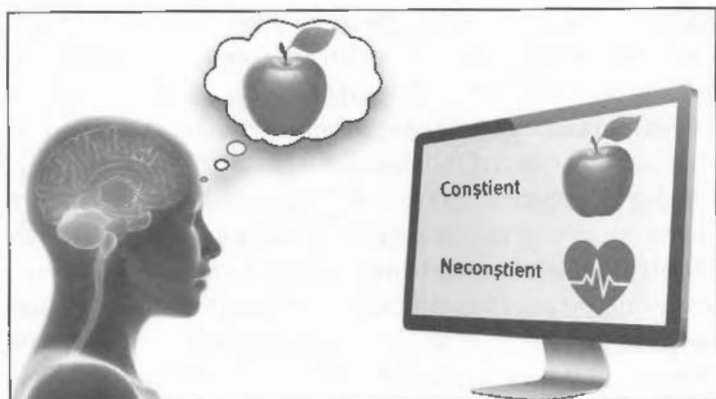


Figura 8.2 Să presupunem că un calculator măsoară informația procesată în creierul tău și prezice de care părți din ea ești conștient, conform unei teorii despre conștiință. Poți testa științific teoria verificând dacă predicțiile ei sunt corecte prin comparație cu experiența ta subiectivă.

Pe de altă parte, dacă teoria refuză să facă vreo predicție, răspunzând mereu „incert”, ea este netestabilă, deci neștiințifică. Asta s-ar putea întâmpla fiindcă e aplicabilă doar în anumite situații, fiindcă sunt prea greu de efectuat calculele necesare ori fiindcă senzorii cerebrali nu sunt buni. Teoriile științifice curente se află de regulă undeva la mijloc, oferind răspunsuri testabile la unele întrebări, dar nu la toate. De pildă, teoria noastră centrală din fizică nu răspunde la întrebări despre sisteme care sunt în același timp extrem de mici (implicând mecanica cuantică) și extrem de grele (implicând relativitatea generală), fiindcă n-am găsit încă ecuațiile matematice pentru acest caz. Această teorie centrală nu prezice nici masele exacte ale tuturor atomilor – în acest caz credem că avem ecuațiile necesare, însă n-am reușit deocamdată să le calculăm cu precizie soluțiile. Cu cât o teorie trăiește mai periculos asumându-și riscuri și făcând predicții testabile, cu atât e mai utilă și o luăm mai în serios dacă supraviețuiește tuturor încercărilor noastre de a o ucide. Într-adevăr, putem testa doar *unele* predicții ale teoriilor despre conștiință, dar asta e valabil pentru *toate* teoriile din fizică. Prin urmare, în loc să pierdem vremea lamentându-ne

pentru ceea ce nu putem testa, n-avem decât să ne punem pe treabă testând ce *putem* testa.

Pe scurt, orice teorie ce prezice care sisteme fizice sunt conștiente (problema destul de grea) este științifică atâta vreme cât poate prezice care dintre procesele tale cerebrale sunt conștiente. Chestiunea testabilității devine însă mai puțin clară pentru întrebările de rang mai înalt în ierarhia din figura 8.1. Ce-ar însemna ca o teorie să prezică felul în care percepi subiectiv culoarea roșie? Iar dacă o teorie pretinde că explică de ce există conștiința, cum o testezi experimental? N-ar trebui să evităm asemenea întrebări doar fiindcă sunt dificile – ne vom întoarce la ele, dar cred că, atunci când ne confruntăm cu mai multe întrebări similare rămase fără răspuns, e înțelept să începem prin a o aborda pe cea mai simplă. De aceea cercetările mele de la MIT privind conștiința sunt concentrate direct asupra bazei piramidei din figura 8.1. Am discutat recent această strategie cu colegul meu fizician Piet Hut de la Princeton, care mi-a spus în glumă că a încerca să construiești vârful piramidei înainte să construiești baza e ca și cum te-ai fi preocupat interpretarea mecanicii cuantice înainte de descoperirea ecuației lui Schrödinger, fundamentul matematic care ne permite să prezicem rezultatele experimentelor noastre.

Când vorbim despre ce se află dincolo de știință, e important să ținem cont de faptul că răspunsul depinde de epocă! Cu patru secole în urmă, Galileo Galilei a fost atât de impresionat de teoriile fizice întemeiate pe matematică, încât a spus că natura este „o carte scrisă în limbajul matematicii”. Dacă arunca un strugure și o alună, putea prezice cu exactitate forma traiectoriilor lor și momentul când ating solul. Nu avea însă nici un indiciu pentru a explica de ce strugurele e galben, iar alună maronie, sau de ce unul e moale, iar cealaltă dură – aceste aspecte ale lumii se aflau dincolo de domeniul științei la acea vreme. Dar nu pentru totdeauna! Când James Clerk Maxwell a descoperit în 1861 ecuațiile care-i poartă numele, a devenit limpede că lumina și culorile pot fi înțelese matematic. Știm acum că ecuația lui Schrödinger pomenită mai sus, descoperită în 1925, poate fi folosită pentru a prezice

toate proprietățile materiei, inclusiv ce e moale sau dur. Dacă progresul teoretic ne-a permis să facem tot mai multe predicții științifice, progresul tehnologic ne-a permis să facem tot mai multe teste experimentale: aproape tot ce studiem acum cu telescoape, microscopice sau acceleratoare de particule s-a aflat odinioară în afara științei. Cu alte cuvinte, domeniul de competență al științei s-a extins spectaculos începând cu epoca lui Galilei, de la o parte minusculă din ansamblul fenomenelor la un procentaj mare, incluzând particulele subatomice, găurile negre și originile noastre din urmă cu 13,8 miliarde de ani. Se pune atunci întrebarea: ce rămâne?

După părerea mea, conștiința este elefantul din cameră. Nu doar știi că ești conștient, dar asta e *tot* ce știi cu deplină certitudine – restul e deducție, după cum a arătat René Descartes, contemporanul lui Galilei. Progresul teoretic și tehnologic va așeza oare ferm în cele din urmă conștiința în domeniul științei? Nu știm, așa cum nici Galilei nu știa dacă într-o bună zi vom înțelege lumina și materia.* Un singur lucru e cert: nu vom reuși dacă nu încercăm! De aceea eu împreună cu alți oameni de știință din lumea întreagă ne străduim să formulăm și să testăm teorii despre conștiință.

INDICII EXPERIMENTALE PRIVIND CONȘTIINȚA

Procesarea informației are loc în capetele noastre chiar acum. Cât din aceasta e conștientă, și cât nu e? Înainte de a examina teoriile despre conștiință și ce prezic ele, să vedem ce am aflat până în prezent din experimente, de la observații tradiționale folosind o

* Dacă realitatea noastră fizică e în întregime de ordin matematic (simplu spus, e bazată pe informație), după cum am explorat în cartea mea *Our Mathematical Universe*, atunci nici un aspect al realității – și nici măcar conștiința – nu se află în afara domeniului științei. Problema cu adevărat grea a conștiinței este, din această perspectivă, exact aceeași problemă ca aceea de a înțelege felul în care ceva de natură matematică poate simți ceea ce e fizic: dacă o parte a unei structuri matematice e conștientă, atunci va percepe restul ca pe lumea fizică externă. (N. a.)

tehnologie rudimentară sau fără să apeleze deloc la tehnologie, până la măsurători cerebrale de ultimă oră.

Ce comportamente sunt conștiente?

Dacă înmulțești mental 32 cu 17, ești conștient de multe din procesele interioare ale calculului. Dar să presupunem că îți arăt un portret al lui Albert Einstein și te invit să-mi spui ce reprezintă. După cum am văzut în capitolul 2, și aceasta e o sarcină computațională: creierul tău evaluează o funcție al cărei input e informația primită de la ochi despre un număr mare de pixeli, iar al cărei output e informația trimisă către mușchii care controlează gura ta și corzile vocale. Informaticienii numesc această sarcină „clasificarea imaginilor“, urmată de „sinteza vorbirii“. Deși acest calcul e incomparabil mai complicat decât sarcina de a înmulți două numere, îl efectuezi mult mai repede, aparent fără efort și fără să fii conștient de *felul* în care îl faci. Experiența ta subiectivă constă doar în a privi imaginea, a avea senzația că recunoști și a te auzi pe tine însuși rostind „Einstein“.

Psihologii știu de multă vreme că poți executa neconștient o gamă largă de alte sarcini și poți manifesta o gamă largă de alte comportamente, de la reflexul de a clipi la respirație, întinderea mâinii, apucare și păstrarea echilibrului. De regulă, ești conștient de ce ai făcut, dar nu și de cum ai făcut. Pe de altă parte, comportamentele care implică situații neobișnuite, autocontrolul, reguli logice complicate, raționamentul abstract sau folosirea limbajului sunt în general conștiente. Ele se numesc *corelativele comportamentale ale conștiinței* (*behavioral correlates of consciousness*), și sunt strâns legate de modul lent și controlat de a gândi, cerând mult efort, cunoscut în psihologie drept „Sistemul 2“.⁵

Se știe de asemenea că multe rutine conștiente pot deveni neconștiente prin exercițiu intens, cum ar fi mersul, înotul, mersul pe bicicletă, șofatul, dactilografiatul, bărbieritul, legarea șireturilor, jocul pe calculator și cântatul la pian.⁶ Într-adevăr, e bine cunoscut faptul că experții își îndeplinesc cel mai bine sarcinile specifice atunci când se află într-o stare „fluidă“, conștienți doar

de ce se întâmplă la un nivel înalt și ignorând detaliile de nivel inferior privind felul în care fac ceea ce fac. De pildă, încearcă să citești următoarea frază fiind conștient de fiecare literă în parte, ca și cum abia înveți să citești. Nu simți că procesul e mult mai lent decât în cazul în care ești doar conștient de text la nivelul cuvintelor și ideilor?

Procesarea neconștientă a informației nu numai că e posibilă, dar pare să fie mai curând regula decât excepția. Dovezile arată că din cei aproximativ 10^7 biți de informație care ajung în creierul nostru în fiecare secundă de la organele senzoriale, nu putem fi conștienți decât de o parte minusculă, estimările variind de la 10 la 50 de biți.⁷ Aceasta sugerează că procesarea informației de care suntem conștienți reprezintă doar vârful aisbergului.

Luate împreună, aceste indicii i-au făcut pe unii cercetători să susțină că procesarea conștientă a informației este un fel de director executiv al minții noastre, care se ocupă doar de cele mai importante decizii presupunând analiza complexă a datelor din întregul creier.⁸ Asta ar explica de ce, la fel ca directorul executiv al unei companii, de regulă nu vrea să fie distrasă cunoscând tot ce fac subordonații – dar poate afla, dacă dorește. Pentru a percepe această atenție selectivă în acțiune, privește din nou cuvântul „directorul”: fixează-ți privirea asupra punctului din „i” și, fără să-ți miști ochii, deplasează-ți atenția de la punct asupra întregii litere, apoi asupra întregului cuvânt. Deși informația provenită de la retină a rămas aceeași, experiența ta conștientă s-a schimbat. Comparația cu directorul executiv explică și de ce dexteritatea devine neconștientă: după ce s-a chinuit să deprindă cititul și dactilografierea, directorul executiv delegă aceste sarcini de rutină subordonaților neconștienți, pentru a se putea concentra asupra unor noi provocări de nivel înalt.

Unde se află conștiința?

Experimente și analize ingenioase au sugerat că nu doar la anumite comportamente se limitează conștiința, dar și la anumite părți din creier. Care sunt principalii suspecti? Majoritatea primelor indicii

au provenit de la pacienți cu leziuni cerebrale: vătămări ale creierului localizate, provocate de accidente, atacuri cerebrale, tumori sau infecții. Rezultatele nu erau însă întotdeauna concludente. De pildă, faptul că leziuni din regiunea dorsală a creierului pot provoca orbirea înseamnă oare că acolo e sediul văzului conștient, sau înseamnă pur și simplu că informația vizuală trece pe-acolo în drum spre ceea ce va deveni conștient, la fel cum trece mai întâi prin ochi?

Deși leziunile și intervențiile medicale n-au localizat cu precizie experiențele conștiente, ele au îngustat sfera posibilităților. De exemplu, știu că deși resimt durerea în mână ca și cum acolo s-ar ivi, senzația dureroasă trebuie să apară în alt loc, fiindcă un chirurg mi-a suprimat durerea din mână fără să facă nimic mâinii: mi-a anesteziat nervii din umăr. Mai mult, unii oameni cărora li s-a amputat mâna simt o durere „fantomă” ca și cum ar apărea în mâna lor inexistentă. Un alt exemplu: am observat odată că, atunci când priveam doar cu ochiul meu drept, lipsea o parte din câmpul meu vizual – un doctor a determinat că retina începea să se dezlipească și mi-a lipit-o la loc. Dimpotrivă, pacienți cu anumite leziuni cerebrale suferă de *neglijare hemispațială*: și lor le lipsește informația din jumătate din câmpul vizual, dar nu-și dau seama de asta – de pildă, incapabili să observe restul, mănâncă doar ce se află în jumătatea stângă a farfuriei. E ca și cum conștiința privind jumătate din lumea lor ar fi dispărut. Generează însă oare acele zone vătămate din creier experiența spațială, sau doar oferă informații sediilor conștiinței, așa cum făcea retina mea?

În anii 1930, neurochirurgul canadian de origine americană Wilder Penfield a constatat că pacienții îi spuneau că diferite părți ale corpului lor sunt atinse atunci când stimula electric anumite zone din ceea ce se numește acum *cortexul somatosenzorial* (figura 8.3).⁹ A constatat de asemenea că aceștia își mișcau involuntar anumite părți ale corpului atunci când stimula zone din ceea ce se numește acum *cortexul motor*. Înseamnă însă asta oare că procesarea informației în aceste zone cerebrale corespunde atingerii și mișcării conștiente?

Din fericire, tehnologia modernă ne oferă indicii mult mai amănunțite. Deși suntem încă departe de a măsura fiecare activare a tuturor celor aproximativ o sută de miliarde de neuroni, tehnologia de citire a creierului progresează rapid, implicând tehnici cu denumiri intimidante precum IRMf, EEG, MEG, ECoG, eFiz și detecție pe bază de fluorescență. Imagistica prin rezonanță magnetică funcțională (IRMf) măsoară proprietățile magnetice ale nucleelor de hidrogen pentru a alcătui o hartă 3-D a creierului în fiecare secundă, la o rezoluție de ordinul milimetrului. Electroencefalografia (EEG) și magnetoencefalografia (MEG) măsoară câmpul electric și magnetic din jurul capului pentru a cartografia creierul de mii de ori pe secundă, dar cu o rezoluție mai slabă, fără a putea distinge detalii mai fine de câțiva centimetri. Dacă ești temător, te vei bucura să afli că toate aceste trei tehnici sunt neinvazive. Dacă însă nu te deranjează să-ți fie deschisă cutia crani-ană, există și alte alegeri. Electrocuticografia (ECoG) presupune plasarea a vreo sută de cabluri electrice pe suprafața creierului tău, în vreme ce electrofiziologia (eFiz) implică introducerea unor microconductorți (adesea mai subțiri decât firul de păr) adânc în creier, pentru a înregistra simultan tensiuni din o sută de locuri. Numeroși pacienți cu epilepsie își petrec zile întregi în spital în vreme ce ECoG e folosită pentru a descoperi ce parte a creierului declanșează crizele și trebuie rezecată, și acceptă în același timp ca neurologii să facă asupra lor experimente privind conștiința. În fine, detecția pe bază de fluorescență presupune manipularea genetică a neuronilor pentru a emite scânteii atunci când se activează, ceea ce permite măsurarea activității lor cu ajutorul microscopului. Dintre toate tehnicile, aceasta poate monitoriza rapid cel mai mare număr de neuroni, cel puțin la animale cu creier transparent – cum sunt viermele *C. elegans*, având 302 neuroni, și larva peștelui-zebră, cu circa 100 000 de neuroni.

Deși Francis Crick l-a sfătuit pe Christof Koch să nu studieze conștiința, Christof n-a vrut să renunțe, iar în cele din urmă l-a atras de partea lui și pe Francis. În 1990 ei au scris împreună un celebru articol despre ceea ce numeau „corelativele neuronale ale

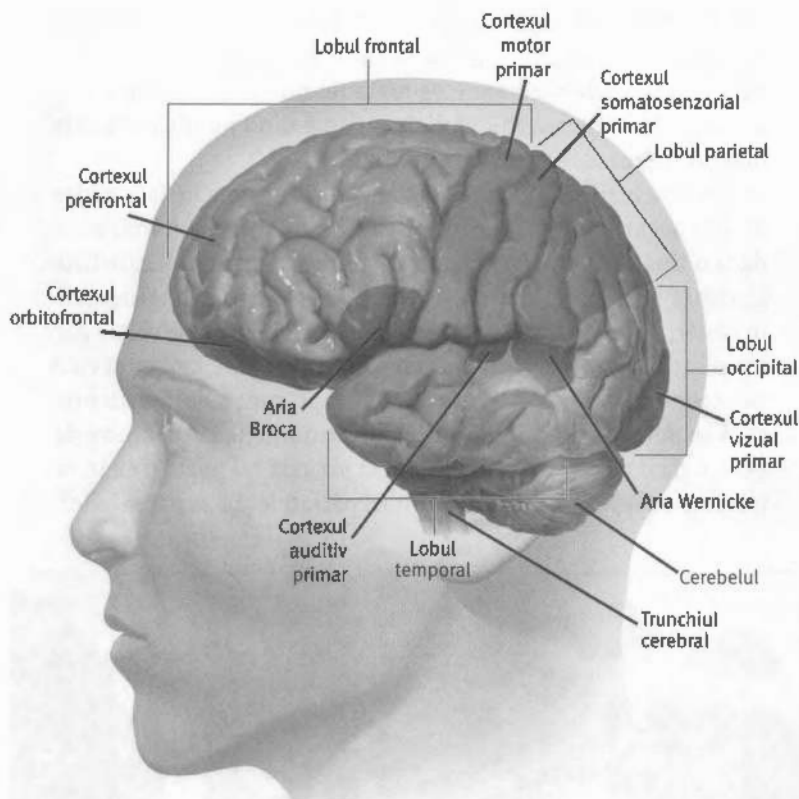


Figura 8.3 Cortexurile vizual, auditiv, somatosenzorial și motor sunt implicate în văz, auz, simțul tactil și, respectiv, activarea mișcării – ceea ce nu dovedește însă că acolo își are sediul *conștiința* văzului, auzului, atingerii și mișcării. Conform unor experimente recente, cortexul vizual primar e complet neconștient, împreună cu cerebelul și trunchiul cerebral. Imagine obținută prin amabilitatea agenției Lachina (www.lachina.com).

conștiinței“ (*neural correlates of consciousness – NCC*), în care își puneau problema să determine care anume procese cerebrale corespund experiențelor conștiente. Vreme de mii de ani, gânditorii avuseseră acces la procesarea informației din creierul lor doar prin experiența subiectivă și comportament. Crick și Koch au arătat că tehnologia de citire a creierului oferea brusc un acces

independent la această informație, permițând studiul științific al corespondenței dintre procesarea informației și experiența conștientă. Măsurătorile bazate pe noua tehnologie au adus în centrul atenției căutarea NCC, mii de articole fiind publicate în cele mai prestigioase reviste de neurologie.¹⁰

Ce concluzii se pot trage deocamdată? Pentru a ne face o idee despre această muncă detectivistică, să ne întrebăm mai întâi dacă retina ta e conștientă, sau e doar un sistem zombi care înregistrează informația vizuală, o procesează și o trimite mai departe în creier, unde are loc experiența vizuală subiectivă. Privește partea stângă a figurii 8.4: care pătrat e mai întunecat, cel notat cu A sau cel notat cu B? A, nu-i așa? Nu, de fapt, sunt colorate identic, ceea ce poți verifica uitându-te la ele prin micile fante dintre degete. Asta dovedește că experiența ta vizuală nu poate ține în întregime de rețină, fiindcă altminteri pătratele ți s-ar părea la fel.

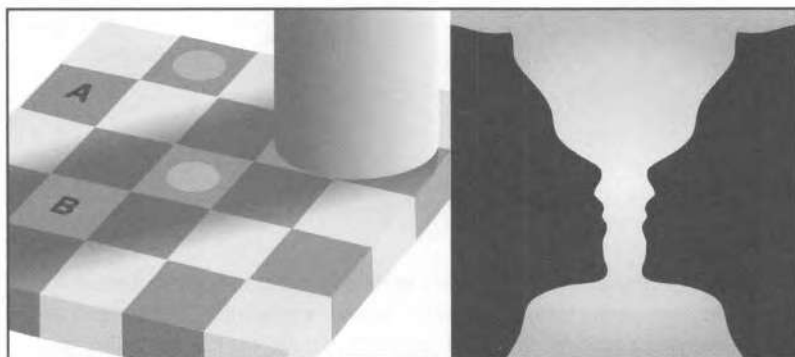


Figura 8.4 Care pătrat e mai întunecat, A sau B? Ce vezi în dreapta, o cupă, două femei sau ambele succesiv? Iluzii optice precum acestea demonstrează că nu în ochi sau în alte prime elemente care apar în procesul vederii se poate afla conștiința ta vizuală, fiindcă ea nu depinde doar de ceea ce este în imagine.

Privește acum partea dreaptă a figurii 8.4. Vezi două femei sau o cupă? Dacă te uiți suficient de mult timp, le vei vedea pe amândouă succesiv, deși informația care ajunge la retina ta rămâne aceeași. Măsurând ce se întâmplă în creier în cele două situații,

putem găsi de unde provine diferența – și nu-i vorba de retină, care se comportă identic în ambele cazuri.

Lovitura fatală dată ipotezei că retina ar fi conștientă vine de la o tehnică numită „suprimarea continuă a instantaneelor” (*continuous flash suppression*) introdusă de Christof Koch, Stanislas Dehaene și colaboratorii lor: s-a descoperit că dacă unul din ochi urmărește un șir complicat de tipare care se modifică rapid, aceasta va distra sistemul tău vizual în asemenea măsură încât nu vei sesiza deloc o imagine statică prezentată celuilalt ochi.¹¹ Pe scurt, poți avea o imagine vizuală în retina ta fără s-o percepi, iar (în timp ce visezi) poți percepe o imagine fără ca ea să se afle în retina ta. Asta dovedește că cele două retine nu găzduiesc conștiința ta vizuală, așa cum nici o cameră video n-o găzduiește, deși efectuează calcule complicate implicând peste o sută de milioane de neuroni.

Cercetătorii NCC folosesc de asemenea suprimarea continuă a instantaneelor, iluziile optice/auditive instabile și alte trucuri ca să descopere ce regiuni ale creierului *sunt* responsabile pentru fiecare dintre experiențele conștiente. Strategia de bază e compararea a ceea ce neuronii fac în două situații în care în esență totul (inclusiv input-ul tău senzorial) este identic – cu excepția experienței tale conștiente. Părțile din creier care conform măsurătorilor se comportă diferit sunt apoi identificate drept NCC.

Asemenea cercetări NCC au demonstrat că nimic din conștiința ta nu se află în intestine, deși acolo e localizat sistemul nervos enteric, cu o jumătate de miliard de neuroni care calculează cum să digere optim hrana. Nimic din conștiința ta nu se află în trunchiul cerebral, partea inferioară a creierului care se leagă de măduva spinării și controlează respirația, ritmul cardiac și tensiunea arterială. Încă mai surprinzător, conștiința ta nu pare să se extindă la cerebel (figura 8.3), care conține circa două treimi din toți neuronii tăi: pacienții al căror cerebel e distrus vorbesc nedeslușit și se mișcă stângaci, semănând cu bețivii, dar sunt pe deplin conștienți.

Care părți ale creierului *sunt* responsabile pentru conștiință rămâne o întrebare fără un răspuns cert. Unele cercetări NCC

recente sugerează să sediul conștiinței ar fi într-o „zonă fierbinte” implicând talamusul (aproape de centrul creierului) și partea dorsală a cortexului (învelișul extern al creierului constând într-o suprafață încrețită alcătuită din șase straturi, care, dacă ar fi întinsă, ar avea aria unei fețe de masă).¹² Aceleași cercetări sugerează că o excepție o constituie cortexul vizual primar, situat în extremitatea dorsală a creierului, care e la fel de neconștient ca globii oculari sau retinele.

Când apare conștiința?

Am prezentat până acum indicii privind acele tipuri de procesare a informației care sunt conștiente și acele locuri unde apare conștiința. Dar *când* apare conștiința? În copilărie credeam că devenim conștienți de evenimente pe măsură ce ele se petrec, fără vreun decalaj temporal. Deși am în continuare aceeași senzație subiectivă, e limpede că nu poate fi corectă, din moment ce durează un timp până ce creierul meu procesează informația care intră prin organele mele de simț. Cercetătorii NCC au măsurat acest decalaj, iar Christof Koch susține că trece aproximativ un sfert de secundă între momentul în care lumina intră în ochii tăi provenind de la un obiect complex și momentul în care percepi în mod conștient că îl vezi.¹³ Asta înseamnă că atunci când șofezi pe o autostradă cu viteza de nouăzeci de kilometri pe oră și vezi brusc o vererită la câțiva metri în fața ta, ai călcat-o deja!

Pe scurt, conștiința ta trăiește în trecut, decalajul față de lumea exterioară fiind, după estimarea lui Christof Koch, de aproximativ un sfert de secundă. Și totuși, poți adesea reacționa la lucruri mai repede decât poți deveni conștient de ele, ceea ce dovedește că procesarea informației responsabilă pentru cele mai rapide reacții trebuie să fie neconștientă. De pildă, dacă un obiect străin se apropie de ochii tăi, reflexul de a clipi poate închide pleoapa ta într-o zecime de secundă. Este ca și cum unul dintre sistemele tale cerebrale primește informațiile amenințătoare de la sistemul vizual, calculează că ochiul tău riscă să fie lovit, transmite

mușchilor ochiului instrucțiuni să clipească și, simultan, transmite părții conștiente din creier mesajul „Hei, o să clipim“. În momentul în care acest mesaj e citit și inclus în experiența ta conștientă, ai clipit deja.

Într-adevăr, sistemul care citește acel mesaj e continuu bombardat cu mesaje din întreg corpul tău, unele mai întârziate decât altele. Durează mai mult ca semnalele nervoase să ajungă la creier de la degete decât de la obraz din cauza distanței, și durează mai mult să analizezi imaginile decât sunetele, fiindcă e mai complicat – de aceea la cursele atletice startul se dă printr-o împușcătură, nu printr-un indiciu vizual. Și totuși, atunci când îți atingi nasul, resimți în mod conștient senzația nasului simultan cu senzația vârfului degetului, iar atunci când bați din palme, vezi, auzi și simți bătaia exact în același timp.¹⁴ Asta înseamnă că întreaga ta experiență conștientă privind un eveniment e creată abia după ce au ajuns și cele mai întârziate mesaje și au fost analizate.

Un faimos set de experimente NCC inițiate de fiziologul Benjamin Liber demonstrează că acele acțiuni pe care le poți efectua neconștient nu se limitează la reacții rapide cum sunt clipitul și loviturile din ping pong, ci includ și anumite decizii pe care le-ai pune pe seama liberului-arbitru – măsurătorile cerebrale pot uneori prezice hotărârea ta înainte să devii conștient că ai luat-o.¹⁵

TEORII DESPRE CONȘTIINȚĂ

Am văzut că, deși încă nu înțelegem conștiința, avem mari cantități de date experimentale privind diverse aspecte ale ei. Toate aceste date provin însă de la *creier*, prin urmare ce ne-ar putea spune despre conștiința din *mașini*? Aceasta presupune o extrapolare majoră, dincolo de domeniul nostru experimental curent. Cu alte cuvinte, e nevoie de o teorie.

De ce o teorie?

Pentru a înțelege de ce, să comparăm teoriile despre conștiință cu cele despre gravitație. Oamenii de știință au început să ia în serios teoria newtoniană a gravitației fiindcă au obținut de la ea mai mult decât au introdus în ea: ecuații simple care încăpeau pe un șervețel de hârtie puteau prezice corect rezultatul fiecărui experiment privind gravitația efectuat vreodată. Prin urmare, ei au luat în serios predicțiile ei și mult în afara domeniului în care fusese testată, iar aceste extrapolări îndrăznețe s-au dovedit a funcționa chiar și pentru mișcările galaxiilor din roiuri de galaxii întinse pe milioane de ani-lumină. Predicțiile erau însă ușor incorecte pentru mișcarea lui Mercur în jurul Soarelui. Oamenii de știință au început să ia în serios mai buna teorie einsteiniană a gravitației, relativitatea generală, fiindcă era mai elegantă și mai economică, și făcea predicții corecte și acolo unde teoria lui Newton dădea greș. Prin urmare, au luat în serios și predicțiile ei mult dincolo de domeniul în care fusese testată, pentru fenomene exotice cum sunt găurile negre, undele gravitaționale în însăși țesătura spațiu-timp și expansiunea universului nostru dintr-o origine fierbinte – toate acestea fiind ulterior confirmate experimental.

În mod asemănător, dacă o teorie matematică a conștiinței ale cărei ecuații încap pe un șervețel de hârtie ar prezice cu succes rezultatele tuturor experimentelor efectuate pe creier, am începe să luăm în serios nu doar teoria însăși, dar și predicțiile ei pentru o conștiință aflată în afara creierului – de pildă, într-o mașină.

Conștiința din perspectiva fizicii

Deși unele teorii despre conștiință apar încă din Antichitate, majoritatea celor moderne sunt întemeiate pe neuropsihologie și neuroștiințe, încercând să explice și să prezică conștiința în termenii evenimentelor neuronale care au loc în creier.¹⁶ Cu toate că aceste teorii au făcut unele predicții corecte privind corelativele neuronale ale conștiinței, ele nu pot și nu urmăresc să facă predicții privind conștiința mașinilor. Pentru saltul de la creiere

la mașini, trebuie să generalizăm de la NCC la PCC: corelativele fizice ale conștiinței (*physical correlates of consciousness*), definite ca tipare conștiente ale particulelor în mișcare. Dacă o teorie poate prezice corect ce este și ce nu este conștiința referindu-se doar la constituenți fizici de bază cum sunt particulele elementare și câmpurile de forțe, atunci ea poate face predicții nu numai pentru creiere, dar și pentru orice alte aranjamente ale materiei, inclusiv viitoarele sisteme IA. Să considerăm deci o perspectivă a fizicii: ce aranjamente de particule sunt conștiente?

Ajungem astfel la o altă întrebare: cum poate ceva atât de complex precum conștiința să fie alcătuit din ceva atât de simplu precum particulele? Este vorba, cred, despre un fenomen care are proprietăți aflate deasupra și dincolo de cele ale particulelor. În fizică, asemenea fenomene se numesc „emergente”.¹⁷ Pentru a le înțelege, vă propun să analizăm un fenomen emergent mai simplu decât conștiința: umiditatea.

O picătură de apă e umedă, însă un cristal de gheață și un nor de aburi nu sunt, deși au în compoziție molecule de apă identice. De ce? Fiindcă proprietatea de a fi umed depinde doar de aranjamentul moleculelor. N-are sens să spui că o singură moleculă de apă e umedă, pentru că fenomenul umidității apare numai dacă există multe molecule, aranjate în tiparul pe care îl numim lichid. Prin urmare, solidele, lichidele și gazele sunt, toate, fenomene emergente: ele reprezintă mai mult decât suma componentelor, fiindcă au proprietăți aflate deasupra și dincolo de proprietățile componentelor lor. Posedă proprietăți pe care componentele lor nu le au.

La fel ca în cazul solidelor, lichidelor și gazelor, conștiința cred că e un fenomen emergent, cu proprietăți aflate deasupra și dincolo de cele ale particulelor care o alcătuiesc. De pildă, intrarea într-un somn profund abrogă conștiința pur și simplu prin rearanjarea particulelor. La fel, conștiința mea ar dispărea dacă aș muri înghețat, ceea ce mi-ar rearanja particulele într-un mod mult mai nefericit.

Când pui multe particule laolaltă pentru a alcătui orice, de la apă la creier, apar noi fenomene cu proprietăți observabile. Nouă,

fizicienilor, ne place să studiem aceste proprietăți emergente, care pot fi adesea identificate printr-un mic set de numere măsurabile – mărimi precum vâscozitatea unei substanțe, compresibilitatea ei etc. De pildă, dacă o substanță e atât de vâscoasă încât e rigidă, o numim solidă, altminteri o numim fluidă. Dacă un fluid e incompresibil, îl numim lichid, altminteri îl numim gaz sau plasmă, în funcție de conductivitatea lui electrică.

Conștiința ca informație

Pot exista așadar mărimi analoge care să cuantifice conștiința? Specialistul italian în neuroștiințe Giulio Tononi a propus o asemenea mărime, pe care o numește *informație integrată*, notată cu litera grecească Φ (Phi) și care măsoară în esență cât de mult știu una despre alta diferite părți ale unui sistem (vezi figura 8.5).

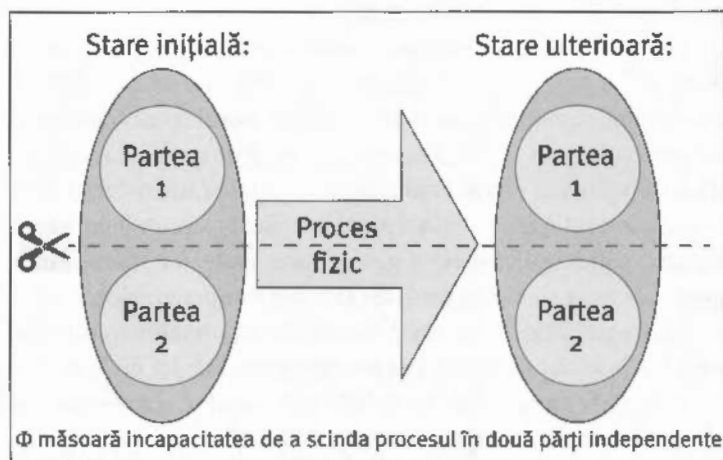


Figura 8.5 Pentru un proces fizic care, odată cu trecerea timpului, transformă starea inițială a unui sistem într-o nouă stare, *informația integrată* Φ măsoară incapacitatea de a scinda procesul în două părți independente. Dacă starea viitoare a fiecărei părți depinde doar de propriul ei trecut, și nu de ceea ce a făcut cealaltă parte, atunci $\Phi = 0$: ceea ce numeam un sistem reprezintă de fapt două sisteme independente care nu comunică unul cu altul.

L-am cunoscut pe Giulio în 2014 la o conferință de fizică din Puerto Rico, unde i-am invitat pe el și pe Christof Koch, și mi-a părut a fi renașcentistul perfect care ar fi putut sta alături de Galilei și Leonardo da Vinci. Modestia sa nu putea ascunde incredibilele lui cunoștințe artistice, literare și filozofice, iar reputația sa culinară mi-a fost adeverită de un jurnalist TV care mi-a spus recent că în doar câteva minute Giulio i-a pregătit cea mai delicioasă salată pe care a mâncat-o vreodată. Mi-am dat curând seama că în spatele exprimării sale cumpănite se ascunde un intelect cutezător, gata să urmeze dovezile oriunde l-ar aduce, indiferent față de ideile preconcepute și de tabuurile dominante. La fel cum Galilei și-a construit teoria matematică a mișcării în ciuda interdicției de a pune în discuție geocentrismul, Giulio a elaborat cea mai precisă matematică dintre teoriile despre conștiință de care dispunem, *teoria informației integrate* (*integrated information theory* – IIT).

Am susținut vreme de decenii că de fapt conștiința este ceea ce simte informația atunci când e procesată în anumite moduri complexe.¹⁸ IIT îmi dă dreptate și înlocuiește exprimarea mea vagă „anumite moduri complexe” cu o definiție precisă: procesarea informației trebuie să fie integrată, adică Φ trebuie să fie mare. Argumentul lui Giulio este pe cât de puternic, pe atât de simplu: sistemul conștient trebuie să fie integrat într-un ansamblu unificat, fiindcă dacă ar fi constituit din două părți independente, s-ar simți ca două entități conștiente separate, nu ca una singură. Cu alte cuvinte, dacă o parte conștientă a unui creier sau a unui calculator nu poate comunica cu restul, atunci restul nu poate fi o parte a experienței subiective a acesteia.

Giulio și colaboratorii săi au măsurat o versiune simplificată a lui Φ folosind EEG pentru a înregistra răspunsul creierului la stimularea magnetică. Acest „detector de conștiință” funcționează foarte bine: a determinat că pacienții erau conștienți când erau treji sau visau, dar neconștienți când erau aneșteziați sau în somn profund. A detectat conștiința chiar și la doi pacienți suferind de sindromul *locked-in*,*

* Literal, „sindromul închiderii”, numit și pseudocomă, în care aproape toți mușchii de control voluntar sunt paralizați, iar pacientul, care de altfel

care nu se puteau mișca sau comunica normal.¹⁹ Avem așadar de-a face cu o tehnologie promițătoare pentru ca în viitor medicii să-și dea seama dacă anumiți pacienți sunt conștienți sau nu.

Ancorarea conștiinței în fizică

IIT e definită doar pentru sisteme discrete care se pot afla într-un număr finit de stări, ca biții din memoria unui calculator sau neuronii ultrasimplificați care sunt fie activați, fie dezactivați. Ca atare, din păcate IIT nu e definită pentru majoritatea sistemelor fizice tradiționale, care pot varia în mod continuu – de pildă, poziția unei particule sau intensitatea câmpului magnetic pot lua orice valoare dintr-un spectru infinit.²⁰ Dacă încerci să aplici formula IIT la asemenea sisteme, obții de regulă rezultatul absurd că Φ e infinit. Sistemele cuantice pot fi discrete, dar IIT inițială nu e definită pentru sisteme cuantice. Cum putem deci ancora IIT și alte teorii despre conștiință bazate pe informație pe un fundament fizic solid?

O putem face pornind de la ceea ce am aflat în capitolul 2 despre felul în care grămezi de materie pot avea proprietăți emergente care sunt asociate cu informația. Am văzut că pentru ca ceva să fie folosit ca dispozitiv de memorie care stochează informația, trebuie să aibă mai multe stări cu timp de viață lung. Am văzut de asemenea că pentru ca o substanță să fie *computroniu* (să poată efectua calcule), este nevoie în plus de o dinamică complexă: legile fizicii trebuie să o poată transforma în moduri suficient de complicate pentru a fi efectuată orice procesare a informației. În fine, am văzut că o rețea neuronală, de exemplu, este un puternic substrat pentru învățare, fiindcă, supunându-se pur și simplu legilor fizicii, se poate rearanja pentru a deveni tot mai bine adaptată efectuării calculelor dorite. Se pune acum încă o întrebare: ce face ca o bucată de materie să aibă experiențe subiective? Cu alte cuvinte, în ce condiții o bucată de materie va putea face următoarele patru lucruri?

e perfect conștient, nu poate comunica decât clipind sau eventual mișcându-și globii oculari. (N. t.)

1. memorează
2. calculează
3. învață
4. are experiențe

Despre primele trei am vorbit în capitolul 2, să ne ocupăm acum de al patrulea lucru. La fel cum Margolus și Toffoli au creat termenul *computroniu* pentru o substanță care poate efectua calcule arbitrare, îmi place să folosesc termenul *sentroniu* pentru cea mai generală substanță care are experiențe subiective (simte).*

Dar cum de este resimțită conștiința ca un fenomen non-fizic, dacă de fapt e un fenomen fizic? Cum se poate simți independentă de substratul ei fizic? Eu cred că *este* independentă de substratul fizic, materia din care ea alcătuiește un tipar! Am întâlnit multe exemple frumoase de structuri independente de substrat în capitolul 2, între care undele, memoriile și calculele. Am văzut că acestea nu erau doar mai mult decât părțile lor componente (adică erau emergente), dar și independente de părțile lor, având propria lor viață. De pildă, am văzut că o minte simulată sau un personaj dintr-un joc pe computer nu și-ar putea da seama dacă sunt rulate pe Windows, Mac OS, pe un telefon Android sau pe vreun alt sistem de operare, fiindcă ar fi independente de substrat. N-ar putea ști nici dacă porțile logice ale calculatorului sunt alcătuite din tranzistori, circuite optice sau alt hardware. Și nici care sunt legile fundamentale ale fizicii – acestea pot fi oricum, cu condiția să permită construirea calculatoarelor universale.

Pe scurt, conștiința cred că e un fenomen fizic care se simte non-fizic deoarece este la fel ca undele și calculele: are proprietăți independente de substratul ei fizic. Asta decurge logic din ideea de conștiință-ca-informație. Ajungem așadar la o idee radicală, care îmi place foarte mult: dacă conștiința este ceea ce simte informația

* Deși am folosit în trecut „perceptroniu” ca sinonim pentru sentroniu, această denumire sugerează o definiție prea îngustă, fiindcă percepțiile sunt doar acele experiențe subiective pe care le resimțim pe baza input-ului senzorial – excluzând, de pildă, visele și gândurile generate intern. (N. a.)

când e procesată în anumite moduri, atunci trebuie să fie independentă de substrat; contează doar structura procesării informației, nu și structura materiei care efectuează procesarea informației. Cu alte cuvinte, conștiința este de două ori independentă de substrat!

După cum am văzut, fizica descrie tipare în spațiu-timp corespunzând particulelor ce se mișcă de jur-împrejur. Dacă aranjamentele particulelor se supun anumitor principii, ele dau naștere fenomenelor emergente care sunt în mare măsură independente de substratul particulelor și le resimt pe acestea într-un mod total diferit. Un exemplu este procesarea informației în computroniu. Am dus însă acum această idee la un alt nivel: *dacă însăși procesarea informației se supune anumitor principii, ea poate da naștere unui fenomen emergent de nivel mai înalt pe care îl numim conștiință*. Asta așază experiența ta conștientă nu cu un nivel, ci cu două niveluri deasupra materiei. Nu-i de mirare că mintea ta se simte non-fizică!

Se pune deci următoarea întrebare: care sunt principiile cărorora trebuie să li se supună procesarea informației pentru a fi conștientă? Nu pretind că știu ce condiții sunt *suficiente* pentru a garanta conștiința, dar iată patru condiții *necesare* pe care aș paria și pe care le-am explorat în cercetările mele:

Principiu	Definiție
Principiul informației	Un sistem conștient are o mare capacitate de stocare a informației.
Principiul dinamicii	Un sistem conștient are o mare capacitate de procesare a informației.
Principiul independenței	Un sistem conștient are o mare independență față de restul lumii.
Principiul integrării	Un sistem conștient nu poate fi alcătuit din părți aproape independente.

După cum am spus, conștiința cred că este ceea ce simte informația atunci când e procesată în anumite moduri. Pentru a fi așadar conștient, un sistem fizic trebuie să poată stoca și procesa informație – de aici rezultă primele două principii. Observați că memoria nu e necesar să fie de lungă durată: vă recomand să priviți

filmulețul cu Clive Wearing, care pare perfect conștient, deși amintirile lui durează mai puțin de un minut.²¹ Cred că un sistem conștient trebuie de asemenea să fie relativ independent de restul lumii, fiindcă altminteri n-ar simți în mod subiectiv că are o existență independentă. În fine, cred că un sistem conștient trebuie să fie integrat într-un ansamblu unificat, după cum susținea Giulio Tononi, fiindcă dacă ar fi constituit din două părți independente, s-ar simți ca două entități conștiente separate, nu ca una singură. Primele trei principii implică *autonomia*: sistemul e capabil să rețină și să proceseze informație fără să interfereze prea mult cu exteriorul, determinându-și astfel propriul viitor. Luate împreună, cele patru principii arată că sistemul e autonom, dar nu și părțile lui.

Dacă aceste patru principii sunt corecte, lucrurile se limpezesc: trebuie să căutăm teorii riguroase matematic care să le încorporeze, și să testăm experimental teoriile. Trebuie de asemenea să vedem dacă e nevoie și de alte principii. Indiferent dacă IIT este corectă sau nu, cercetătorii ar trebui să elaboreze teorii concurente și să testeze toate teoriile disponibile prin experimente mai bune.

COTROVERSE PRIVIND CONȘTIINȚA

Am vorbit deja despre eterna dispută în jurul întrebării dacă cercetările legate de conștiință sunt un nonsens neștiințific și o pierdere de vreme. Recent au apărut controverse în cercetările de vârf privind conștiința – vă propun să le examinăm pe cele mai interesante dintre acestea.

Teoria informației integrate a lui Tononi a fost întâmpinată nu doar cu laude, dar și cu critici, unele chiar nimicitoare. Iată ce scria nu de mult Scott Aaronson pe blogul său: „După părerea mea, faptul că teoria informației integrate e falsă – demonstrabil falsă, din motive care se află în miezul ei – o așază în ceva de genul top 2% dintre toate teoriile matematice despre conștiință propuse vreodată. Cred că aproape toate teoriile concurente despre conștiință au fost atât de vagi, de superficiale și de maleabile, încât pot

doar aspira la falsitate.²² Spre cinstea lor, Scott și Giulio s-au purtat absolut civilizat în dezbaterile despre IIT de la Universitatea din New York, fiecare ascultând politicos argumentele celuilalt. Aaronson a arătat că în anumite rețele simple de porți logice informația e puternic integrată (Φ mare), iar din moment ce în mod evident acestea nu ar fi conștiente, IIT e falsă. Giulio a ripostat afirmând că, dacă aceste rețele ar fi construite, ele *ar fi* conștiente, iar presupunerea contrară a lui Scott era marcată de antropocentrism, cam la fel ca afirmația unui proprietar de abator cum că animalele nu pot fi conștiente fiindcă nu vorbesc și sunt foarte diferite de oameni. În opinia mea, cu care ambii au fost de acord, dezacordul lor ținea de faptul că pentru Scott integrarea era doar o condiție *necesară*, în vreme ce pentru Giulio era și o condiție *suficientă*. Această din urmă afirmație e mai tranșantă, și sper ca în curând să poată fi testată experimental.²³

Altă afirmație IIT controversată este că arhitecturile calculatoarelor de azi nu pot fi conștiente, fiindcă felul în care se conectează porțile lor logice oferă o integrare foarte scăzută.²⁴ Cu alte cuvinte, dacă te uploadezi într-un robot viitor de mare putere care simulează cu precizie fiecare dintre neuronii și sinapsele tale, atunci, chiar dacă această clonă digitală arată, vorbește și acționează într-un fel prin care să nu poată fi deosebită de tine, Giulio susține că va fi un zombi neconștient fără experiențe subiective – o dezamăgire pentru tine, dacă te-ai uploadat în căutarea nemuririi subiective.* Această afirmație a fost contrazisă de David Chalmers și de profesorul de IA Murray Shanahan, ei imaginându-și ce s-ar întâmpla dacă ți-ar fi treptat înlocuite circuitele neuronale din creier cu un ipotetic hardware digital care le simulează perfect.²⁵ Deși comportamentul tău n-ar fi afectat de înlocuire, din moment ce simularea e presupusă a fi perfectă, *experiența* ta ar trebui să se modifice – din conștientă, la început, în neconștientă, la sfârșit –, dacă

* Există o tensiune potențială între această afirmație și ideea de independență de substrat a conștiinței, din moment ce, deși procesarea informației poate fi diferită la nivelul cel mai de jos, ea este prin definiție identică la nivelurile superioare, acolo unde determină comportamentul. (N. a.)

ar fi să-i dăm crezare lui Giulio. Dar ce ai simți în etapele intermediare, pe măsură ce tot mai multe circuite neuronale sunt înlocuite? Când părțile din creier responsabile de experiența conștientă a jumătății superioare a câmpului tău vizual ar fi înlocuite, ai observa că jumătate din peisajul vizual a dispărut brusc, dar că în chip misterios știi totuși ce se află acolo, așa cum se întâmplă cu pacienții suferind de „vedere oarbă”^{*26} Ar fi într-adevăr tulburător, pentru că, dacă ai simți în mod conștient vreo diferență, atunci le-ai putea vorbi despre asta și prietenilor tăi – însă am presupus că nu apare nici o modificare în comportamentul tău. Singura posibilitate logică, în acord cu acele presupuneri, este că, exact în momentul când oricare dintre lucruri dispăre din conștiința ta, mintea ta e alterată în chip misterios, așa încât fie minți și negi că experiența ta s-a modificat, fie uiți că lucrurile fuseseră diferite.

Pe de altă parte, Murray Shanahan recunoaște că aceeași critică legată de înlocuirea treptată poate fi aplicată *oricărei* teorii care afirmă că poți acționa conștient fără să fii conștient, așa încât ai putea conchide că a acționa conștient și a fi conștient sunt unul și același lucru, iar comportamentul observabil din exterior este deci tot ce contează. Atunci însă cazi în capcana de a prezice că ești neconștient în timp ce visezi, deși știi că nu-i așa.

A treia controversă privind IIT se leagă de întrebarea dacă o entitate conștientă poate fi alcătuită din părți care sunt conștiente în mod separat. De pildă, poate societatea în ansamblul ei să capete conștiință fără ca oamenii să și-o piardă pe a lor? Poate un creier conștient să fie alcătuit din părți care să aibă și ele propria lor conștiință? Conform IIT, răspunsul este „nu”, dar nu toată lumea e convinsă. De pildă, unii pacienți cu leziuni grave care le reduc comunicarea dintre cele două emisfere cerebrale resimt „sindromul mâinii străine”, în care emisfera dreaptă comandă mâinii stângi să facă lucruri pe care pacienții afirmă că nu le provoacă sau nu le înțeleg – uneori ajung chiar să-și folosească cealaltă mână pentru a-și opri mâna „străină”. Cum putem fi siguri că nu există două

* Capacitatea pacienților cu leziuni corticale de a reacționa la stimuli vizuali pe care nu-i văd în mod conștient. (N. t.)

conștiințe separate în capul lor, una în emisfera dreaptă, incapabilă să vorbească, cealaltă în emisfera stângă, responsabilă de întreaga comunicare verbală și afirmând că vorbește în numele ambelor emisfere? Imaginați-vă că prin tehnologia viitorului s-ar realiza comunicarea directă între două creiere umane, iar treptat ar crește capacitatea acestei legături până când comunicarea ar fi la fel de eficientă între creiere ca în interiorul lor. Este oare posibil ca la un moment dat cele două conștiințe individuale să dispară brusc și să fie înlocuite de una singură, unificată, așa cum prezice IIT, sau tranziția ar fi treptată, așa încât conștiințele individuale să coexiste sub o anumită formă chiar și atunci când ar începe să apară experiența unificată?

Altă controversă fascinantă se leagă de întrebarea dacă nu cumva experimentele subestimează totalitatea lucrurilor de care suntem conștienți. Am văzut mai devreme că deși *simțim* că suntem vizual conștienți de o imensă cantitate de informații despre culori, forme, obiecte și practic tot ce se află în fața noastră, experimentele au arătat că putem ține minte și evoca doar o infimă parte din acestea.²⁷ Unii cercetători au încercat să lămurească această discrepanță întrebându-se dacă nu cumva uneori avem „conștiință fără acces”, adică experiența subiectivă a lucrurilor care sunt prea complexe pentru a intra în memoria noastră de lucru spre a fi utilizate mai târziu.²⁸ De pildă, atunci când ajungi la *orbire din neatenție*, fiind prea distras ca să observi un obiect aflat chiar în fața ta, asta nu înseamnă că n-ai experiența vizuală a lui, ci doar că ea n-a fost stocată în memoria ta de lucru.²⁹ Nu-i vorba aici mai curând de uitare decât de orbire? Alți cercetători resping ideea că nu poți avea încredere în ce spun oamenii că resimt, și ne avertizează în privința implicațiilor ei. Murray Shanahan își imaginează un test clinic în care pacienții mărturisesc că au scăpat complet de durere printr-un nou medicament miraculos, care e totuși respins de o comisie guvernamentală: „Pacienții doar își închipuie că au scăpat de durere. Grație neuroștiințelor, noi știm mai bine.”³⁰ Pe de altă parte, au existat cazuri în care pacienților care s-au trezit din întâmplare în timpul operației li s-a administrat un

medicament ca să uite calvarul prin care au trecut. Ar trebui să ne încredem în declarațiile lor ulterioare cum că n-au simțit nici o durere?³¹

CE AR PUTEA SIMȚI O IA CONȘTIENȚĂ?

Dacă un sistem IA din viitor va fi conștient, care va fi experiența lui subiectivă? Aceasta e esența problemei „și mai grea” a conștiinței, iar astfel ajungem la al doilea nivel de dificultate din figura 8.1. Nu numai că deocamdată ne lipsește o teorie care să răspundă la această întrebare, dar nici măcar nu suntem siguri că e logic posibil să răspunzi pe deplin la ea. În fond, cum ar suna un răspuns satisfăcător? Cum să-i explici unui om născut orb cum arată culoarea roșie?

Din fericire, incapacitatea noastră actuală de a da un răspuns complet nu ne împiedică să dăm răspunsuri parțiale. Extraterestri inteligenți care ar studia sistemul senzorial uman ar deduce probabil despre culori că sunt qualia percepute a fi asociate cu fiecare punct de pe o suprafață bidimensională (câmpul nostru vizual), în vreme ce sunetele nu sunt percepute a fi localizate spațial, iar durerile sunt percepute a fi asociate cu diferite părți ale corpului nostru. Din descoperirea faptului că retinele noastre au trei tipuri de celule conice fotosensibile, ar putea deduce că percepem trei culori primare, iar toate celelalte qualia de culoare rezultă din combinarea acestora. Măsurând cât timp durează ca neuronii să transmită informația prin creier, ar putea trage concluzia că avem aproximativ zece gânduri sau percepții conștiente pe secundă, iar atunci când privim la televizor filme cu 24 de cadre pe secundă, nu le percepem ca pe un șir de imagini statice, ci ca pe o mișcare continuă. Măsurând cât de rapid e eliberată adrenalina în sânge și cât timp trece până dispare, ar putea deduce că accesele noastre de furie se declanșează în câteva secunde și durează câteva minute.

Aplicând raționamente similare bazate pe fizică, putem lansa unele ipoteze privind felul în care ar putea simți o conștiință

artificială. În primul rând, spațiul experiențelor IA posibile e *urias* în comparație cu ceea ce putem resimți noi, oamenii. Avem câte o clasă de qualia pentru fiecare dintre simțurile noastre, dar o IA poate avea incomparabil mai multe tipuri de senzori și de reprezentări interne ale informației, așa încât trebuie să evităm capcana de a presupune că o IA s-ar simți la fel ca un om.

În al doilea rând, o conștiință artificială de dimensiunea creierului nostru ar avea de milioane de ori mai multe experiențe pe secundă decât noi, fiindcă semnalele electromagnetice se propagă cu viteza luminii – de milioane de ori mai repede decât semnalele neuronale. Însă cu cât o IA este mai mare, cu atât trebuie să fie mai lente gândurile ei globale, pentru a da timp informației să circule între toate părțile sale, după cum am văzut în capitolul 4. Este așadar de așteptat ca o IA „Geea” de dimensiunea Pământului să aibă doar vreo zece experiențe conștiente pe secundă, la fel ca un om, iar o IA de dimensiunea unei galaxii ar avea doar un gând global cam la fiecare 100 000 de ani – adică nu mai mult de vreo 100 000 de gânduri în întreaga istorie a universului nostru! Asta ar face ca o IA mare să-și delege calculele celor mai mici subsisteme capabile să le efectueze, la fel cum mintea noastră conștientă a delegat reflexul de a clipi unui subsistem mic, rapid și neconștient. Am văzut că procesarea conștientă a informației în creierul nostru pare să fie doar vârful unui aisberg neconștient, dar este de așteptat să se ajungă și mai departe pentru marile IA din viitor: dacă au o conștiință unică, atunci e probabil să ignore aproape întreaga procesare a informației care are loc în ele. În plus, deși experiențele conștiente de care se bucură pot fi extrem de complexe, acestea se desfășoară în ritm de melc în comparație cu activitățile rapide ale părților mai mici.

Aceasta accentuează controversa legată de întrebarea dacă părțile unei entități conștiente pot fi și ele conștiente. IIT prezice că nu, ceea ce înseamnă că dacă o IA astronomic de mare din viitor e conștientă, atunci aproape întreaga procesare a informației e neconștientă. Prin urmare, dacă o civilizație de IA mici își îmbunătățește capacitățile de comunicare până în punctul în care

apare o unică minte agregată conștientă, atunci conștiințele ei individuale mult mai rapide dispar brusc. Pe de altă parte, dacă predicția IIT e falsă, mintea agregată poate coexista cu tot restul minților conștiente mai mici. Ne putem chiar închipui o ierarhie de conștiințe la toate nivelurile, de la microscopic la cosmic.

După cum am văzut mai sus, procesarea neconștientă a informației în creierele noastre umane este legată de un mod de gândire rapid, automat, care nu necesită efort, numit de psihologi „Sistemul 1”.³² De pildă, Sistemul 1 ar putea da de știre conștiinței tale că extrem de complexa lui analiză asupra datelor input-ului vizual a dezvăluit că prietenul tău cel mai bun tocmai a sosit, fără să-ți dea vreun indiciu despre felul în care a efectuat calculul. Dacă această legătură între sisteme și conștiință se dovedește a fi valabilă, va fi tentant să generalizăm terminologia pentru IA, numind toate sarcinile rapide de rutină delegate subunităților neconștiente Sistemul 1 al IA. Gândirea globală lentă, controlată și necesitând efort, presupunând că e conștientă, ar fi Sistemul 2 al IA. Noi, oamenii, avem experiențe conștiente implicând și ceea ce voi numi „Sistemul 0”: percepția pasivă brută care are loc chiar și atunci când stai fără să te miști sau să te gândești, și nu faci decât să observi lumea din jur. Complexitatea Sistemelor 0, 1 și 2 pare să crească progresiv, așa încât e frapant faptul că doar cel din mijloc e neconștient. IIT explică asta spunând că informația senzorială brută din Sistemul 0 e stocată în structuri cerebrale de tip rețea, cu un grad înalt de integrare, în vreme ce Sistemul 2 are un grad înalt de integrare datorită buclelor de feedback, în care toate informațiile de care ești conștient acum pot influența stările viitoare ale creierului tău. Pe de altă parte, tocmai predicția privind rețeaua conștientă a declanșat critica mai sus pomenită a lui Scott Aaronson la adresa IIT. Pe scurt, dacă o teorie care rezolvă problema destul de grea a conștiinței va putea trece într-o bună zi o serie de teste experimentale riguroase, așa încât să-i luăm în serios predicțiile, atunci se va îngusta mult spectrul opțiunilor pentru problema și mai grea legată de ce ar putea simți o IA conștientă din viitor.

Unele aspecte ale experienței noastre subiective au în mod evident origini evolutive – de pildă, dorințele noastre emoționale legate de autoconservare (a mânca, a bea, a evita să fii ucis) și de reproducere. Ar fi deci posibil să creăm o IA care să nu resimtă niciodată *qualia* precum foamea, setea sau dorința sexuală. După cum am văzut în capitolul precedent, dacă o IA extrem de inteligentă e programată să aibă practic orice obiectiv, oricât de ambițios, s-ar strădui probabil să se autoconserve pentru a-l îndeplini. Dacă ar exista însă o societate de inteligențe artificiale, acestora le-ar putea lipsi puternica noastră teamă de moarte: câtă vreme au copii ale lor însele, tot ce ar putea pierde ar fi amintirile acumulate de la ultimul backup, dacă au încredere că software-ul copiat va fi utilizat. În plus, capacitatea inteligențelor artificiale de a copia cu ușurință informație și software le-ar reduce probabil acel individualism caracteristic conștiinței umane: diferența dintre tine și mine ar scădea dacă ne-am putea partaja și copia amintirile și capacitățile, așa încât un grup de inteligențe artificiale învecinate s-ar simți mai curând ca un unic organism cu o minte agregată.

Ar simți oare o conștiință artificială că posedă liber-arbitru? Observați că deși filozofii și-au pus de mii de ani problema liberului-arbitru legat de *noi*, nu s-a ajuns la un consens nici măcar în privința felului în care trebuie pusă întrebarea.³³ Eu îmi pun o altă problemă, pesemne mai ușor de abordat. Voi încerca să vă conving că răspunsul e pur și simplu: „Da, orice entitate care ia decizii conștiente va *simți* subiectiv că are liber-arbitru, indiferent dacă e biologică sau artificială.” Deciziile se întind pe un spectru având aceste extreme:

1. Știi exact de ce ai făcut o anumită alegere.
2. Habar n-ai de ce ai făcut o anumită alegere – simți că ai ales la întâmplare sau după capriciu.

Discuțiile despre liberul-arbitru sunt de regulă concentrate în jurul încercării de a reconcilia comportamentul orientat spre obiective al celui care ia decizii cu legile fizicii. Dacă ai de ales între următoarele două explicații pentru ce ai făcut, care e cea

corectă? „Am rugat-o să ne întâlnim, fiindcă îmi place foarte mult” sau „Particulele care mă alcătuiesc au făcut-o, mișcându-se conform legilor fizicii”. În capitolul precedent am văzut însă că *ambele* sunt corecte: ceea ce simți că e comportament orientat spre obiective poate apărea din legile deterministe ale fizicii, lipsite de vreun obiectiv. Mai precis, când un sistem (creier sau IA) ia o decizie de tipul 1, el calculează ce să hotărască folosind un algoritm determinist, iar motivul pentru care simte că el a hotărât este acela că într-adevăr el a luat hotărârea atunci când a calculat ce să facă. Mai mult, după cum a subliniat Seth Lloyd,³⁴ există în informatică o celebră teoremă care spune că pentru aproape toate calculele nu există un mod mai rapid de a determina rezultatul lor decât de a le face efectiv. Asta înseamnă că în general ți-e imposibil să-ți dai seama, într-un interval de timp mai mic de o secundă, ce vei hotărî peste o secundă să faci, ceea ce îți întărește sentimentul că ai liber-arbitru. Dimpotrivă, când un sistem (creier sau IA) ia o decizie de tipul 2, el își programează pur și simplu mintea să se bazeze în hotărâre pe rezultatul unui subsistem care acționează ca un generator de numere aleatorii. În creiere și în calculatoare, numerele aleatorii sunt cu ușurință generate prin amplificarea zgomotului. Indiferent unde se află o decizie în spectrul de la 1 la 2, atât conștiințele biologice, cât și cele artificiale simt așadar că au liber-arbitru: simt că într-adevăr ele sunt cele care hotărăsc și nu pot prezice cu certitudine ce decizie vor lua până n-au încheiat procesul de gândire.

Unii oameni îmi spun că găsesc cauzalitatea degradantă, că privează de sens procesul gândirii și îi reduce la statutul de „simple” mașini. Asemenea negativism mi se pare absurd și neîntemeiat. În primul rând, nu e nimic „simplu” legat de creierul omului, care, după părerea mea, este cel mai complex obiect fizic din universul cunoscut. În al doilea rând, ce ar prefera ei? N-ar vrea ca propriile lor procese de gândire (calculele efectuate de creier) să fie acelea care iau hotărârile lor? Experiența subiectivă a liberului-arbitru arată cum sunt resimțite calculele lor din interior: ei nu cunosc rezultatul calculului până nu-l încheie. Calculul *este* decizia.

Să ne întoarcem acum în punctul de pornire al acestei cărți: cum vrem să fie viitorul vieții? Am văzut în capitolul precedent că toate culturile își doresc un viitor plin de experiențe pozitive, dar nu ajungem la consens în privința a ceea ce e pozitiv și a compromisului între ce e bine pentru diferite forme de viață. Asemenea controverse nu trebuie totuși să ne abată atenția de la elefantul din cameră: nu pot exista experiențe pozitive dacă nu există deloc experiențe – dacă nu există conștiință. Cu alte cuvinte, fără conștiință nu pot exista fericire, bine, frumos, sens sau rost, ci doar o risipă astronomică de spațiu. Prin urmare, atunci când oamenii își pun întrebări despre sensul vieții ca și cum ar fi sarcina cosmosului să dea sens existenței noastre, ei privesc lucrurile invers: *nu universul nostru dă sens ființelor conștiente, ci ființele conștiente dau sens universului nostru*. Principalul nostru obiectiv pentru viitor ar trebui să fie menținerea (și, să sperăm, extinderea) conștiinței biologice și/sau artificiale în cosmosul nostru, în loc s-o ducem la pieire.

Dacă reușim, ce vor simți oamenii față de coexistența cu mașini tot mai deștepte? Vădita ascensiune inexorabilă a inteligenței artificiale te neliniștește – și dacă da, de ce? În capitolul 3 am văzut că ar fi relativ ușor ca o tehnologie folosind IA să ne satisfacă nevoi de bază cum ar fi securitatea și bunăstarea materială, atâta vreme cât există voința politică pentru asta. Poate însă că a fi bine hrănit, îmbrăcat, adăpostit și distrat ți se pare insuficient. Dacă ni se garantează că IA se va ocupa de toate dorințele și nevoile noastre practice, nu vom sfârși oare prin a simți că viața noastră e lipsită de sens și de scop, că vom trăi la fel ca animalele bine îngrijite de la grădina zoologică?

Tradițional, noi, oamenii, ne-am întemeiat orgoliul pe ideea de *exceptionalism uman*: convingerea că suntem entitățile cele mai deștepte de pe planetă, așadar unici și superiori. Ascensiunea IA ne va obliga să devenim mai modești. Ideea de superioritate față de alții (indivizi, grupuri etnice, specii etc.) a provocat mari nenorociri, și a sosit pesemne momentul să renunțăm la ea. Nu numai

că excepționalismul uman e responsabil de suferințe din trecut, dar el nici nu-i necesar pentru înflorirea umanității: dacă am descoperi o civilizație extraterestră pașnică mult mai avansată decât noi în știință, artă și toate celelalte lucruri care ne interesează, asta nu i-ar împiedica pe oameni să simtă că viața lor are un sens și un scop. Am rămâne cu familiile, prietenii și comunitățile mai largi, cu toate activitățile care ne dau un sens și un scop, pierzând, să sperăm, doar aroganța.

Din moment ce vorbim despre viitor, să ne gândim nu numai la sensul vieților noastre, dar și la sensul universului nostru. În această privință, doi dintre fizicienii pe care îi admir cel mai mult, Steven Weinberg și Freeman Dyson, au vederi diametral opuse. Weinberg, care a primit Premiul Nobel pentru contribuții fundamentale la modelul standard din fizica particulelor, a spus: „Cu cât universul pare mai comprehensibil, cu atât pare mai lipsit de sens.”³⁵ Pe de altă parte, după cum am văzut în capitolul 6, Dyson e mult mai optimist: deși este de acord că universul *a fost* lipsit de sens, el crede că viața îi oferă acum tot mai mult sens, iar viitorul va arăta și mai bine dacă viața va reuși să se răspândească prin cosmos. Dyson își încheie astfel celebrul articol din 1979: „Care univers e mai aproape de adevăr, al lui Weinberg sau al meu? Într-o bună zi, nu peste multă vreme, vom afla.”³⁶ Dacă universul nostru va cădea într-o stare de eternă neconștiință pentru că vom distruge viața de pe Pământ sau vom permite unei IA zombi să preia controlul asupra universului nostru, atunci Weinberg va avea câștig de cauză.

Privind din această perspectivă, ne dăm seama că, deși ne-am concentrat în cartea de față asupra viitorului inteligenței, viitorul conștiinței e și mai important, fiindcă el oferă un sens. Filozofii apelează la latină pentru a pune în antiteză *sapientia* (capacitatea de a gândi inteligent) și *sententia* (capacitatea de a percepe subiectiv qualia). Noi, oamenii, ne-am construit identitatea pornind de la faptul că suntem *Homo sapiens*, cele mai deștepte entități de prin părțile noastre. Cum suntem pe cale de a fi întrecuți de mașini și mai deștepte, propun să ne redefinim ca *Homo sentiens*!

REZUMAT:

- Nu există o definiție indiscutabilă a „conștiinței”. Eu folosesc termenul într-un sens larg, neantropocentric: *conștiință = experiență subiectivă*.
- Dacă inteligențele artificiale sunt conștiente în acest sens este ceea ce contează pentru cele mai spinoase probleme etice și filozofice puse de ascensiunea IA: Pot suferi inteligențele artificiale? Ar trebui ele să aibă drepturi? Este uploadarea o sinucidere subiectivă? Ar putea fi un cosmos din viitor plin de IA apocalipsa zombi absolută?
- Problema înțelegerii inteligenței n-ar trebui să fie amestecată cu trei probleme separate legate de conștiință: „problema destul de grea” de a prezice care sisteme fizice sunt conștiente, „problema și mai grea” de a prezice qualia și „problema cu adevărat grea” – de ce ajunge ceva să fie conștient.
- „Problema destul de grea” a conștiinței face obiectul științei, din moment ce o teorie care prezice care dintre procesele tale cerebrale sunt conștiente e testabilă experimental și falsificabilă, însă deocamdată nu e limpede cum ar putea știința rezolva pe deplin cele două probleme mai grele.
- Experimentele neuroștiințifice sugerează că multe comportamente și regiuni ale creierului sunt neconștiente, o mare parte din experiența noastră conștientă reprezentând un rezumat post factum al unei mult mai vaste cantități de informație neconștientă.
- Generalizarea predicțiilor privind conștiința de la creiere la mașini necesită o teorie. Conștiința pare să aibă nevoie nu de un tip anume de particule sau de câmpuri, ci de un tip anume de procesare a informației, în mare măsură autonom și integrat, așa încât sistemul în ansamblu să fie relativ autonom, dar nu și părțile lui.
- Conștiința se simte pesemne atât de non-fizică pentru că este de două ori independentă de substrat: dacă ea este felul în care simte informația atunci când este procesată în anumite moduri complexe, ceea ce contează este doar structura procesării informației, nu și structura materiei care efectuează procesarea informației.
- Dacă e posibilă conștiința artificială, atunci spațiul tuturor experiențelor posibile ale IA este probabil uriaș în comparație cu cel al experiențelor umane, întins pe un spectru vast de qualia și scări temporale – toate împărțind sentimentul că posedă liber-arbitru.
- Din moment ce ne putem pierde orice sens în absența conștiinței, nu universul nostru dă sens ființelor conștiente, ci ființele conștiente dau sens universului nostru.
- Aceasta sugerează că, pe măsură ce ne pregătim să fim întrecuți de mașini tot mai deștepte, ne consolează mai curând faptul că suntem *Homo sentiens*, decât că suntem *Homo sapiens*.

EPILOG

Povestea Echipei FLI

Cel mai trist lucru legat de viața din prezent este că știința acumulează cunoaștere mai rapid decât acumulează societatea înțelepciune. – Isaac Asimov

Iată-ne, dragă cititorule, ajunși la sfârșitul cărții, după ce am explorat originea și destinul inteligenței, obiectivelor și sensului. Cum am putea așadar transforma aceste idei în acțiune? Ce ar trebui să facem concret pentru a ne făuri un viitor cât mai bun? Este întrebarea pe care mi-o pun chiar acum, pe 9 ianuarie 2017, când mă aflu pe scaunul de lângă fereastră, în drumul de întoarcere de la San Francisco la Boston, după conferința despre IA pe care am organizat-o la Asilomar, așa că permiteți-mi să închei această carte împărtășindu-vă gândurile mele.

Meia doarme alături de mine, după ce a petrecut multe nopți cu pregătirile și organizarea. Ce săptămână agitată am avut! Am reușit să-i aducem laolaltă pe aproape toți cei menționați în cartea de față, pentru câteva zile, la această continuare a conferinței din Puerto Rico – între ei, oameni de afaceri ca Elon Musk și Larry Page, cercetători IA de vârf din universități și companii cum sunt DeepMind, Google, Facebook, Apple, IBM, Microsoft și Baidu, precum și economiști, juriști, filozofi și alți gânditori remarcabili (vezi figura 9.1). Rezultatele au depășit până și speranțele mele cele mai cutezătoare, și sunt atât de optimist în privința viitorului vieții cum n-am mai fost de mult. În acest epilog am să explic de ce.

De la vârsta de 14 ani, când am aflat de cursa înarmărilor nucleare, mă îngrijorează ideea că puterea tehnologiei noastre crește mai rapid decât înțelegciunea cu care o controlăm. M-am hotărât așadar să introduc un capitol despre această provocare în prima mea carte, *Our Mathematical Universe*, deși restul ei se ocupa în principal de fizică. În 2014, de Anul Nou, am luat hotărârea că nu mai am voie să deplâng această situație fără să reflectez serios la ce pot face eu însumi în privința asta, și am respectat legământul în cursul turneului de promovare a cărții mele din ianuarie: Meia și cu mine ne-am gândit mult cum putem înființa un soi de organizație nonprofit concentrată pe ameliorarea viitorului vieții prin intermediul tehnologiei.

Ea a insistat să-i dăm un nume pozitiv, în nici un caz ceva de genul „Institutul Apocalipsei și Întinericului” sau „Institutul Hai-să-ne-temem-de-viitor”. Din moment ce Institutul pentru Viitorul Omenirii (Future of Humanity Institute) exista deja, am ales numele de Institutul pentru Viitorul Vieții (Future of Life Institute – FLI), care avea avantajul de a fi mai cuprinzător. Pe 22 ianuarie turneul ne-a adus la Santa Cruz, iar pe când soarele californian apunea deasupra Pacificului, am luat cina cu vechiul nostru prieten Anthony Aguirre și l-am convins să ni se alăture. Nu e numai unul dintre cei mai inteligenți și mai idealiști oameni pe care îi cunosc, dar el a reușit să conducă, împreună cu mine, timp de un deceniu și o altă organizație nonprofit, Institutul pentru Întrebările Fundamentale (Foundational Questions Institute – vezi <http://fqxi.org>).

În săptămâna următoare, turneul de promovare a cărții m-a adus la Londra. Fiindcă mă preocupa viitorul IA, l-am căutat pe Demis Hassabis, care a avut amabilitatea să mă invite să vizitez sediul central al DeepMind. Am fost uimit să descopăr cât se dezvoltase de când Demis mă vizitase la MIT, cu doi ani în urmă, când Google tocmai cumpărase DeepMind cu 650 de milioane de dolari. Văzând acum birourile încăpătoare pline de minți strălucite

care urmăreau obiectivul îndrăzneț al lui Demis de a „rezolva problema inteligenței”, am avut sentimentul că succesul era într-adevăr posibil.

A doua zi de dimineață am vorbit cu prietenul meu Jaan Tallinn prin Skype, software-ul la crearea căruia contribuise. I-am explicat cum vedem noi proiectul FLI, iar peste o oră a hotărât să ne finanțeze cu 100 000 de dolari pe an! M-a impresionat că cineva poate avea atâta încredere în mine, mai ales că un an mai târziu, după conferința din Puerto Rico menționată în capitolul 1, mi-a spus, răsând, că a fost cea mai bună investiție din câte a făcut vreodată.

A doua zi, editorul meu mi-a lăsat o pauză în program, de care am profitat vizitând Muzeul Științei din Londra. Fiind de atâta timp obsedat de trecutul și viitorul inteligenței, am simțit brusc că mă plimbam prin concretizarea fizică a gândurilor mele. Este adunată acolo o colecție fantastică de exponate reprezentând sporirea cunoașterii umane, de la locomotiva Rocket a lui Stephenson la Modelul T al lui Ford, o replică în mărime naturală a modulului lunar al lui Apollo 11 și computere de la calculatorul mecanic al lui Babbage „Mecanismul Diferențial” până la cele din zilele noastre. Este prezentată și evoluția înțelegerii minții, de la experimentele lui Galvani cu picioare de broască până la neuroni, EEG și IRMf.

Îmi vine foarte rar să plâng, dar asta am făcut când am ieșit din muzeu – într-un tunel aglomerat, în drum spre stația de metrou South Kensington. Cei din jur își vedeau fiecare de drumul lui, din fericire fără să bănuiască la ce mă gândeam. La început, noi, oamenii, am descoperit cum să reproducem anumite procese naturale cu mașini, generând vânt, lumină și cai-putere pe cale mecanică. Treptat, am înțeles că și trupurile noastre sunt mașini. Apoi, descoperirea celulei nervoase a început să estompeze granița dintre trup și minte. Am construit mașini care întrec nu numai mușchii, dar și mințile noastre. Și atunci, în paralel cu descoperirea a ceea ce suntem, nu cumva în mod inevitabil devenim desueți? Ar fi tragic.

Acest gând m-a înspăimântat, dar mi-a și întărit hotărârea de a mă ține de legământul făcut de Anul Nou. Simțeam că, pentru



Figura 9.1 Conferința noastră Asilomar din ianuarie 2017, continuarea celei din Puerto Rico, a adus laolaltă un remarcabil grup de cercetători din IA și domenii înrudite. Rândul din spate, de la stânga la dreapta: Patrick Lin, Daniel Weld, Ariel Conn, Nancy Chang, Tom Mitchell, Ray Kurzweil, Daniel Dewey, Margaret Boden, Peter Norvig, Nick Hay, Moshe Vardi, Scott Siskind, Nick Bostrom, Francesca Rossi, Shane Legg, Manuela Veloso, David Marble, Katja Grace, Irakli Beridze, Marty Tenenbaum, Gill Pratt, Martin Rees, Joshua Greene, Matt Scherer, Angela Kane, Amara Angelica, Jeff Mohr, Mustafa Suleyman, Steve Omohundro, Kate Crawford, Vitalik Buterin, Yutaka Matsuo, Stefano Ermon, Michael Wellman, Bas Steunebrink, Wendell Wallach, Allan Dafoe, Toby Ord, Thomas Dietterich, Daniel Kahneman, Dario Amodei, Eric Drexler, Tomaso Poggio, Eric Schmidt, Pedro Ortega, David Leake, Seán Ó hÉigeartaigh, Owain Evans, Jaan Tallinn, Anca Dragan, Sean Legassick, Toby Walsh, Peter Asaro, Kay Firth-Butterfield, Philip Sabes, Paul Merolla, Bart Selman, Tucker Davey, ?, Jacob Steinhardt, Moshe Looks, Josh Tenenbaum, Tom Gruber, Andrew Ng, Kareem Ayoub, Craig Calhoun, Percy Liang, Helen Toner, David Chalmers, Richard Sutton, Claudia Passos-Ferreira, János Krámar, William MacAskill, Eliezer Yudkowsky, Brian Ziebart, Huw Price, Carl Shulman, Neil Lawrence, Richard Mallah, Jurgen Schmidhuber, Dileep George, Jonathan Rothberg, Noah Rothberg. Rândul din față: Anthony Aguirre, Sonia Sachs, Lucas Perry, Jeffrey Sachs, Vincent Conitzer, Steve Goose, Victoria Krakovna, Owen Cotton-Barratt, Daniela Rus, Dylan Hadfield-Menell, Verity Harding, Shimon Zilis, Laurent Orseau, Ramana Kumar, Nate Soares, Andrew McAfee, Jack Clark, Anna Salamon, Long Ouyang, Andrew Critch, Paul Christiano, Yoshua Bengio, David Sanford, Catherine Olsson, Jessica Taylor, Martina Kunz, Kristinn Thorisson, Stuart Armstrong, Yann LeCun, Alexander Tamas,



Roman Yampolskiy, Marin Soljačić, Lawrence Krauss, Stuart Russell, Eric Brynjolfsson, Ryan Calo, ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, Max Tegmark, Adrian Weller, Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt. Au sosit mai târziu: Guru Banavar, Demis Hassabis, Rao Kam-bhampati, Elon Musk, Larry Page, Anthony Romero

a completa echipa de fondatori FLI, mai aveam nevoie de cineva care să conducă un grup de tineri voluntari idealişti. Alegerea firească a fost Viktoriya Krakovna, o strălucită absolventă de Harvard care, pe lângă faptul că fusese medaliată cu argint la Olimpiada Internațională de Matematică, înființase Citadela, unde se aflau o duzină de tineri idealişti care voiau să se afirme. Cinci zile mai târziu, Meia și cu mine am invitat-o la noi pentru a-i vorbi despre planul nostru, iar înainte să terminăm de mâncat sushi-ul FLI se născuse.

AVENTURA DIN PUERTO RICO

Așa a început o aventură uimitoare, care continuă și azi. După cum am menționat în capitolul 1, găzduim cu regularitate la noi acasă întâlniri de *brainstorming*, la care participă zeci de studenți, profesori și persoane interesate de subiect din zona Bostonului,



Figura 9.2 Jaan Tallinn, Anthony Aguirre, al dumneavoastră Max Tegmark, Meia Chita-Tegmark și Viktoriya Krakovna sărbătorind cu sushi înființarea FLI pe 23 mai 2014.

ideile cele mai apreciate fiind transformate în proiecte – primul a fost acel editorial despre care am vorbit în capitolul 1, avându-i drept coautori pe Stephen Hawking, Stuart Russell și Frank Wilczek, care a inițiat dezbaterea publică. În paralel cu primii pași pentru înființarea unei noi organizații (angajări, alegerea unui comitet consultativ și lansarea unui site), a avut loc un eveniment inaugural în fața unui numeros public de la MIT, la care Alan Alda a explorat viitorul tehnologiei împreună cu experți de primă mărime.

Ne-am concentrat în restul anului asupra conferinței din Puerto Rico, care, așa cum am menționat în capitolul 1, urmărea să-i atragă pe cei mai importanți cercetători IA din lume într-o dezbatere privind felul în care IA ar putea deveni benefică. Obiectivul nostru era de a deplasa discuția despre siguranța IA de la temeri la acțiune, de la dispute generate de îngrijorare la acordul asupra unor proiecte concrete de cercetare care ar putea maximiza imediat șansele unui rezultat pozitiv. În pregătirea conferinței, am adunat de pretutindeni idei promițătoare despre siguranța IA, primind reacția comunității la lista noastră tot mai bogată. Cu ajutorul lui Stuart Russell și al unui grup de tineri voluntari devotați, între care Daniel Dewey, János Krámar și Richard Mallah, am distilat aceste priorități de cercetare într-un docu-

ment care să fie discutat la conferință.¹ Obținerea unui consens privind faptul că era mult de lucru în domeniul siguranței IA speram să-i încurajeze pe cercetători să înceapă această muncă. Triumful ultim ar fi fost să convingem pe cineva să finanțeze asemenea cercetări, fiindcă până atunci ele nu primiseră practic nici un sprijin din partea agențiilor de finanțare guvernamentale.

Și iată că își face apariția Elon Musk. Pe 2 august, ne-a atras atenția celebrul său mesaj lansat pe tweeter: „Merită citită *Superintelligence* de Bostrom. Trebuie să fim superprecauți cu IA. Potențial mai periculoasă decât armele nucleare.” A aflat despre eforturile noastre, și am vorbit cu el la telefon câteva săptămâni mai târziu. Deși mă simțeam foarte emoționat și intimidat, rezultatul a fost remarcabil: a acceptat să se alăture comitetului consultativ FLI, să participe la conferință și eventual să finanțeze primul program de cercetare în domeniul siguranței IA, care avea să fie anunțat în Puerto Rico. Asta ne-a entuziasmat și ne-a făcut să ne sporim eforturile pentru a organiza o conferință excelentă, a identifica teme de cercetare promițătoare și a obține pentru acestea sprijinul comunității.

Peste două luni am ajuns în fine să-l întâlnesc pe Elon, sosit la MIT pentru un simpozion. M-am simțit ciudat să mă aflu singur cu el într-o încăpere mică, la doar câteva momente după ce fermecase o mie de studenți de la MIT ca un star rock, însă apoi nu m-am putut gândi decât la proiectul nostru comun. L-am plăcut din prima clipă. Radia sinceritate și m-a impresionat cât de preocupat era de viitorul pe termen lung al omenirii și cu câtă temeritate își transforma aspirația în acțiune. Dorea ca omenirea să exploreze și să colonizeze universul nostru, așa că a înființat o companie spațială. Dorea energie fără poluare, așa că a înființat o companie care să producă energie solară și una care să fabrice mașini electrice. Înalt, chipeș, bun vorbitor și cunoscând o mulțime de lucruri, nu era de mirare că lumea îl asculta cu sufletul la gură.

Din păcate, acest eveniment de la MIT m-a făcut să-mi dau seama ce rol nefast pot juca uneori mediile de informare. Elon a ținut

timp de o oră un discurs fascinant despre explorarea spațiului, care putea fi un formidabil spectacol TV. La urmă, un student i-a pus o întrebare în afara subiectului despre IA. În răspunsul lui, Elon a spus, între altele, „cu inteligența artificială invocăm demonul“ – *sigurul* lucru pe care majoritatea mediilor de informare l-au preluat, și, în general, scos din context. M-a frapat faptul că mulți jurnaliști făceau în mod inconștient *exact opusul* a ceea ce încercam noi să facem în Puerto Rico. Dacă noi voiam să obținem un consens în rândul comunității, subliniind ce ne unește, mediile de informare aveau interesul să sublinieze divizarea. Cu cât puteau anunța o controversă mai aprigă, cu atât mai mare erau ratigul și veniturile lor. În plus, în vreme ce noi voiam să reunim oameni din întreg spectrul opiniilor, pentru ca măcar să se înțeleagă unul pe altul mai bine, relatările din mediile de informare îi întârâtau pe cei cu opinii diferite pe unii împotriva altora și alimentau neînțelegerile publicând doar citatele cele mai provocatoare scoase din context. De aceea am hotărât să nu permitem accesul jurnaliștilor la întâlnirea din Puerto Rico și am impus „Regula de la Chatham House“*, care interzice participanților să dezvăluie ulterior cine ce a spus.**

Până la urmă, conferința noastră din Puerto Rico s-a dovedit a fi un succes, dar n-a fost simplu. Apropierea termenului-limită ne-a obligat să ne înțelegem eforturile. De pildă, a trebuit să vorbesc la telefon sau prin skype cu mulți cercetători IA pentru a aduna o masă critică de participanți care să-i atragă și pe alții, iar momentele

* Regulă stabilită în 1927 la Institutul Regal pentru Afaceri Interne (al cărui sediu e Chatham House) care permite participanților la o întrunire să folosească informațiile căpătate cu acest prilej, dar n-au voie să dezvăluie identitatea și afilierea persoanelor care au oferit acele informații, și nici identitatea celorlalți participanți. (N. t.)

** Această experiență m-a făcut să mă gândesc mai atent la felul în care ar trebui eu însumi să interpretez știrile. Deși, evident, fusesem conștient că majoritatea informațiilor sunt orientate politic, mi-am dat seama acum că ele pot fi tendențioase chiar și atunci când e vorba de alte subiecte decât politica. (N. a.)

dramatice n-au lipsit – ca atunci când m-am trezit la ora șapte dimineața, pe 27 decembrie, pentru a-l suna pe Elon, aflat în Uruguay, și a-l auzi, printr-o legătură telefonică proastă, că-mi spune: „Nu cred că o să meargă.” Era îngrijorat că un program de cercetare în domeniul siguranței IA ar putea genera un fals sentiment de securitate, permițând cercetătorilor recalcitranți să-și continue munca neabătut, preținzând de formă că-i preocupă siguranța. Apoi însă, deși legătura se întrerupea mereu, am vorbit pe larg despre imensele avantaje care s-ar obține dacă am aduce subiectul în centrul atenției și dacă tot mai mulți cercetători ar lucra în domeniul siguranței IA. După ce convorbirea s-a întrerupt definitiv, mi-a trimis unul dintre cele mai frumoase mesaje pe care le-am primit vreodată: „Am pierdut legătura. Oricum, documentele arată bine. Mă bucur să susțin cercetarea cu \$5M timp de trei ani. Sau poate \$10M?”

Patru zile mai târziu, anul 2015 a început bine pentru Meia și pentru mine – înainte de conferință, am avut răgazul să ne relaxăm și am dansat pe o plajă din Puerto Rico luminată de artificii. Și conferința a avut un început foarte bun: s-a ajuns la un remarcabil consens privind faptul că e nevoie de mai multă cercetare în domeniul siguranței IA, iar prin contribuțiile participanților, acel document despre prioritățile în cercetare la care lucraserăm atât de intens a fost îmbunătățit și a căpătat forma finală. Am făcut să circule scrisoarea deschisă în favoarea cercetărilor pentru siguranța IA, și am fost încântat că l-au semnat aproape toți.

Meia și cu mine am avut o minunată întâlnire în camera noastră de hotel cu Elon, care și-a dat acordul asupra planurilor detaliate pentru programul nostru de burse. Pe Meia au impresionat-o felul simplu și deschis cu care vorbea despre viața lui personală și interesul pe care ni-l arăta. A întrebat cum ne-am cunoscut și a ascultat cu plăcere povestea ei amănunțită. A doua zi am filmat un interviu cu el despre siguranța IA și despre motivele pentru care voia s-o sprijine în tot ce părea să aibă șanse de reușită.²

Punctul culminant al conferinței, anunțul donației lui Elon, era programat pentru duminică 4 ianuarie 2015, la ora șapte seara,

iar eu eram atât de agitat, încât m-am foit în pat toată noaptea dinaintea acelei zile. Apoi, cu un sfert de oră înainte să plecăm spre locul unde trebuia să se desfășoare sesiunea, am primit o lovitură! Asistentul lui Elon ne-a sunat și ne-a spus că s-ar părea că Elon nu poate face anunțul. Meia mi-a mărturisit că nu mă văzuse niciodată atât de nervos și dezamăgit. Elon a apărut în cele din urmă, iar eu, în timp ce stăteam de vorbă cu el, număram secunde până la începerea sesiunii. A explicat că mai erau doar două zile până la lansarea rachetei SpaceX, sperând ca pentru prima dată o navă-dronă să aterizeze cu succes, era deci un moment crucial, drept urmare echipa SpaceX nu voia ca mediile de informare să distragă atenția publicului cu alte știri implicându-l pe el. Anthony Agurre, calm și echilibrat ca de obicei, a subliniat că *nimeni* nu-și dorea distragerea atenției, nici Elon, nici comunitatea IA. Am ajuns cu câteva minute întârziere la sesiunea al cărei moderator eram, dar aveam un plan: nu va fi pomenită nici o sumă, ca să ne asigurăm că anunțul nu e incitant pentru presă, și le voi cere tuturor să respecte regula de la Chatham House păstrând secretul asupra anunțului lui Elon timp de nouă zile dacă racheta lui ajunge la stația spațială, indiferent de reușita sau nereușita aterizării; el a spus că va fi nevoie de mai mult timp dacă racheta explodează la lansare.

Numărătoarea inversă pentru anunț a ajuns în fine la zero. Partenerii mei de discuție se aflau în continuare pe scenă, alături de mine: Eliezer Yudkowsky, Elon Musk, Nick Bostrom, Richard Mallah, Murray Shanahan, Bart Selman, Shane Legg și Vernon Vinge. Treptat, aplauzele publicului s-au stins, dar colegii mei erau mai departe pe scenă, fiindcă le spuseseam să rămână, fără să le explic de ce. Meia avea să-mi spună că pulsul ei o luase razna și strângea pe sub masă mâna Viktoriei Krakovna ca să se calmeze. Zâmbeam, știind că sosise momentul mult așteptat și mult sperat.

Am spus că eram bucuros că se ajunsese la un consens privind necesitatea de a spori cercetările pentru ca IA să fie benefică, și că erau multe direcții de cercetare pe care ar trebui să le urmărim de îndată. S-a vorbit însă în această sesiune și despre riscuri serioase,

am adăugat, așa încât ar fi bine să ne ridicăm moralul înainte de a merge la banchetul conferinței. „Așadar îi dau microfonul lui... Elon Musk!” Am simțit că trăiesc un moment istoric când Elon a luat microfonul și a anunțat că va dona o sumă importantă pentru cercetările în siguranța IA. Nu-i de mirare că au izbucnit aplauze furtunoase. După cum plănuiserăm, n-a menționat suma, dar eu știam că erau zece milioane de dolari, așa cum ne înțeleseserăm.

După conferință, Meia și cu mine am fost în vizită la părinții noștri în Suedia și România, și am urmărit cu emoție la Stockholm, împreună cu tatăl meu, lansarea rachetei. Din păcate, încercarea de aterizare s-a încheiat cu ceea ce Elon numește eufemistic RUD, „rapid unscheduled disassembly” (dezasamblare rapidă neprogramată) – echipa sa avea să obțină o amerizare reușită cincisprezece luni mai târziu.³ Toți sateliții au fost însă lansați cu succes pe orbită, la fel ca programul nostru de burse anunțat de Elon pe tweeter milioanele sale de admiratori.⁴

SIGURANȚA IA AJUNGE ÎN CENTRUL ATENȚIEI

Un obiectiv esențial al conferinței din Puerto Rico fusese acela de a aduce în centrul atenției cercetările din domeniul siguranței IA, și am urmărit cu însuflețire această evoluție în mai multe etape. Mai întâi a fost însăși întrunirea, la care numeroși cercetători s-au bucurat să vadă că, implicându-se în acest domeniu, fac parte dintr-o comunitate în continuă creștere. M-au impresionat încurajările primite de la ei. De pildă, profesorul Bart Selman de la Universitatea Cornell mi-a trimis un e-mail în care spunea: „Vă mărturisesc cu toată sinceritatea că n-am văzut niciodată o întrunire științifică atât de bine organizată, atât de pasionantă și de stimulativă intelectual.”

Următoarea etapă a început pe 11 ianuarie, când Elon a lansat pe tweeter mesajul „Dezvoltatori de vârf din inteligența artificială semnează o scrisoare deschisă prin care susțin necesitatea cercetărilor în siguranța IA”,⁵ atașând link-ul unei pagini cu semnături,

unde au apărut curând peste opt mii de nume, lista conținând multe dintre figurile proeminente din IA. Brusc, a devenit mai greu să susții că cei preocupați de siguranța IA nu știau despre ce vorbesc, fiindcă asta înseamnă acum să spui că marii specialiști în IA nu știau despre ce vorbesc. Scrisoarea deschisă a fost preluată de mediile de informare din întreaga lume, și ne-am dat atunci seama că procedaserăm bine interzicându-le jurnaliștilor accesul la conferință. Deși cuvântul cel mai alarmant din scrisoare era „capcane“, acesta a declanșat în presă titluri precum „Elon Musk și Stephen Hawking semnează o scrisoare deschisă în speranța de a împiedica revolta roboților“, ilustrate de terminatori criminali. Dintre sutele de articole pe care le-am găsit, preferatul nostru a fost unul care își bătea joc de celelalte, spunând că „un titlu care evocă androizi călcând în picioare craniile umane face din tehnologia complexă care poate transforma lumea un spectacol de bălci“. ⁶ Din fericire, au fost și multe articole obiective, iar acum trebuia să facem față torentului de semnături, să le verificăm manual pentru a ne proteja credibilitatea și a elimina nume apărute în batjocură, cum ar fi „HAL 9000“, „Terminator“, „Sarah Janette Connor“ sau „Skynet“. Pentru aceasta și pentru viitoarele scrisori deschise, Viktoriya Krakovna și János Krámar au organizat o brigadă de verificatori voluntari, care-i includea pe Jesse Galef, Eric Gastfriend și Ravathi Vinoth Kumar, lucrând în schimburi, astfel încât atunci când Ravathi se ducea la culcare în India, Eric prelua ștafeta la Boston, și așa mai departe.

A treia etapă a început patru zile mai târziu, când Elon a postat pe tweeter un link la anunțul nostru că dona zece milioane de dolari pentru cercetări în siguranța IA. ⁷ Peste o săptămână am lansat un portal online unde cercetători din întreaga lume puteau face cereri de finanțare. Am putut crea sistemul de cereri atât de repede fiindcă Anthony și cu mine ne ocupaserăm de zece ani cu asemenea competiții pentru burse în fizică. Open Philanthropy Project, o instituție filantropică din California orientată spre finanțări de mare impact, a avut generozitatea să adauge o sumă la cea donată de Elon, așa încât puteam oferi mai multe burse. Nu bănuiam

câte cereri vor fi, din moment ce domeniul era nou, iar până la termenul-limită mai rămânea puțin timp. Răspunsul a fost copleșitor – vreo trei sute de echipe din lumea întreagă cereau aproximativ o sută de milioane de dolari. O comisie de profesori și cercetători IA a analizat propunerile și a ales 37 de echipe câștigătoare, care au fost finanțate pe o perioadă de până la trei ani. Când am anunțat lista câștigătorilor, presa a răspuns pentru prima dată nuanțat, fără imagini cu roboți ucigași. Devenea în fine evident că siguranța IA nu era vorbărie goală, ci o activitate utilă, iar multe mari echipe de cercetare s-au alăturat acestui efort.

A patra etapă s-a desfășurat organic în cursul următorilor doi ani, cu numeroase publicații tehnice și zeci de seminarii despre siguranța IA în întreaga lume, ținute de regulă pe lângă cele mai importante conferințe dedicate inteligenței artificiale. Specialiști perseverenți încercaseră de mulți ani să angajeze comunitatea IA în cercetări legate de siguranță, cu succes limitat, dar acum mișcarea luase cu adevărat avânt. O mare parte din acele publicații au fost finanțate prin programul nostru de burse, iar FLI a depus eforturi pentru a organiza și finanța cât mai multe dintre acele seminarii, însă numărul lor a crescut și grație cercetătorilor IA care și-au dedicat timpul și resursele acestui scop. Ca atare, și mai mulți specialiști au aflat despre cercetările legate de siguranță de la colegii lor, descoperind că, în afară de faptul că e utilă, această activitate poate fi și foarte atractivă, implicând probleme matematice și computaționale interesante.

Desigur, nu toți găsesc atractive ecuațiile complicate. La doi ani după conferința din Puerto Rico, am prefătat conferința noastră Asilomar cu un seminar tehnic la care câștigătorii bursei FLI și-au prezentat rezultatele, pe un ecran mare fiind proiectate diapozitive pline cu simboluri matematice. Moshe Vardi, profesor de IA la Universitatea Rice, a spus în glumă că și-a dat seama că am reușit să impunem ca domeniu de cercetare siguranța IA atunci când întâlnirile au devenit plictisitoare.

Creșterea spectaculoasă a activității legate de siguranța IA nu s-a limitat la mediul academic. Amazon, DeepMind, Facebook,

Google, IBM și Microsoft au lansat un parteneriat industrial pentru IA benefică.⁸ Noi donații importante pentru siguranța IA au permis extinderea cercetărilor la cele mai mari organizații nonprofit înrudite cu a noastră: Institutul de Cercetare pentru Inteligența Mașinilor de la Berkeley, Institutul pentru Viitorul Omenirii de la Oxford și Centrul de Studii privind Riscurile la care Este Expusă Existența de la Cambridge. Donații ulterioare de peste zece milioane de dolari au inițiat și alte eforturi îndreptate spre IA benefică: Centrul Leverhulme pentru Viitorul Inteligenței de la Cambridge, Fundația Gates K&L pentru Etică și Tehnologii Computaționale din Pittsburg și Fondul pentru Etica și Administrarea Inteligenței Artificiale din Miami. În fine, dar nu în ultimul rând, angajându-se să participe cu un miliard de dolari, Elon Musk a intrat în parteneriat cu alți oameni de afaceri pentru a lansa OpenAI, o companie nonprofit din San Francisco având drept obiect inteligența artificială benefică. Cercetarea legată de siguranța IA se impusese definitiv.

În paralel cu cercetarea, au sporit și opiniile exprimate, atât individual, cât și colectiv. Parteneriatul industrial pentru IA și-a publicat principiile fondatoare, iar lungi rapoarte cu liste de recomandări au fost publicate de guvernul Statelor Unite, Universitatea Stanford și IEEE (cea mai mare organizație profesională din lume), împreună cu alte zeci de rapoarte și luări de poziție din diverse surse.⁹

Am urmărit să stimulăm discuțiile semnificative în rândul participanților la Asilomar și să aflăm asupra căror puncte, dacă există vreunele, cade de acord această comunitate eterogenă. Lucas Perry și-a asumat așadar sarcina eroică de a citi toate acele documente pe care le-am găsit și de a extrage opiniile prezentate în ele. Într-un efort-maraton inițiat de Anthony Aguirre și încheiat printr-o serie de lungi teleconferințe, echipa noastră FLI a încercat să grupeze laolaltă opiniile similare și să elimine verbiajul birocratic redundant, pentru a obține o listă de principii succinte, incluzând și opiniile nepublicate, dar semnificative, exprimate în discuții informale și cu alte prilejuri. Dar lista conținea încă ambiguități



Figura 9.3 Grupuri de specialiști reflectând asupra principiilor IA la Asilomar.

sau contradicții, și lăsa loc interpretărilor, așa încât cu o lună înaintea de conferință am trimis-o participanților, primind opiniile și sugestiile lor legate de principii noi sau îmbunătățite. Din răspunsul comunității a rezultat o listă a principiilor semnificativ revizuită, pentru a fi folosită la conferință.

La Asilomar, lista a fost mai departe îmbunătățită în două etape. La început, mici grupuri au discutat principiile care le interesau cel mai mult (figura 9.3), au făcut ajustări de detaliu și au adăugat noi principii și versiuni alternative ale unora vechi. În cele din urmă, am efectuat un sondaj în rândul participanților pentru a determina susținerea de care se bucură fiecare versiune a fiecărui principiu.

Acest proces colectiv a fost deopotrivă exhaustiv și epuizant, Anthony, Meia și cu mine luându-ne din timpul de somn și de masă ca să compilăm tot ce era necesar pentru următoarele etape. A fost însă și palpitant. După asemenea discuții amănunțite, complicate, uneori aprige, obținând o gamă largă de reacții, am fost uimiți de nivelul ridicat al consensului la ultimul sondaj în privința multor principii, unele obținând o susținere de peste 97%. Acest consens ne-a permis să așezăm foarte sus ștacheta exigenței pentru lista finală: am păstrat doar principiile cu care au fost de

acord cel puțin 90% dintre participanți. Deși asta a însemnat ca unele principii întrunind o largă apreciere să fie abandonate în ultimul moment, între care și câteva dintre preferatele mele,¹⁰ a permis ca majoritatea participanților să le aprobe pe toate cu inima împăcată, semnând un document care a circulat printre ei. Iată în continuare rezultatul.

PRINCIPIILE IA DE LA CONFERINȚA ASILOMAR

Inteligența artificială a pus deja la dispoziție instrumente benefice care sunt zilnic folosite de oameni din lumea întreagă. Dezvoltarea ei continuă, călăuzită de următoarele principii, va oferi oamenilor posibilități uimitoare în deceniile și secolele care vor veni.

TEME DE CERCETARE

- §1 **Obiectiv de cercetare:** *Obiectivul cercetării IA trebuie să fie acela de a crea o inteligență benefică, nu una nedirecționată.*
- §2 **Finanțarea cercetării:** *Investițiile în IA trebuie însoțite de finanțarea cercetării care să asigure utilizarea benefică a acesteia, incluzând probleme delicate de informatică, științe economice, științe juridice, etică și sociologie, cum ar fi:*
- (a) *Cum putem face ca sistemele IA din viitor să fie extrem de robuste, așa încât să facă ceea ce vrem, fără erori de funcționare și fără a fi vulnerabile în fața hackerilor?*
 - (b) *Cum ne putem spori prosperitatea prin automatizare, fără a afecta resursele și voința oamenilor?*
 - (c) *Cum putem revizui sistemul nostru juridic pentru a fi mai corect și mai eficient, pentru a păstra ritmul cu IA și a gestiona riscurile asociate cu IA?*
 - (d) *La ce set de valori trebuie aliniată IA și ce statut juridic și etic ar trebui să aibă?*
- §3 **Legătura dintre știință și politică:** *Trebuie să existe un dialog constructiv și sănătos între cercetătorii IA și oamenii politici.*
- §4 **Cultura cercetării:** *Trebuie făurită o cultură a cooperării, încrederii și transparenței în rândul cercetătorilor și dezvoltatorilor IA.*

- §5 Evitarea concurenței neloiale: *Echipele care dezvoltă IA trebuie să coopereze activ pentru a împiedica recurgerea la soluții facile care încalcă normele de siguranță.*

ETICĂ ȘI VALORI

- §6 Siguranță: *Sistemele IA trebuie să fie sigure și protejate în decursul întregului timp de viață al funcționării lor, să poată fi verificate acolo unde sunt aplicabile și unde sunt folosite.*
- §7 Transparență față de eșec: *Dacă un sistem IA provoacă daune, trebuie să se poată stabili de ce.*
- §8 Transparență juridică: *Orice implicare a unui sistem autonom în hotărâri judecătorești trebuie să ofere o explicație satisfăcătoare verificabilă de către o autoritate umană competentă.*
- §9 Responsabilitate: *Proiectanții și constructorii sistemelor IA avansate sunt parte interesată în ce privește implicațiile morale ale folosirii, folosirii greșite și acțiunilor acestora, având responsabilitatea și posibilitatea de a modifica acele implicații.*
- §10 Alinierea valorilor: *Sistemele IA cu un grad înalt de autonomie trebuie proiectate așa încât să poată fi asigurată alinierea obiectivelor și comportamentelor lor cu valorile umane pe tot parcursul funcționării acestora.*
- §11 Valorile umane: *Sistemele IA trebuie proiectate și acționate așa încât să fie compatibile cu idealurile umane legate de demnitate, drepturi, libertăți și diversitate culturală.*
- §12 Confidențialitate personală: *Oamenii trebuie să aibă dreptul de a accesa, gestiona și controla datele pe care le generează, dată fiind capacitatea sistemelor IA de a analiza și utiliza acele date.*
- §13 Libertate și confidențialitate: *Aplicarea IA la datele personale nu trebuie să stirbească libertatea reală sau percepută a oamenilor.*
- §14 Beneficiu colectiv: *De tehnologiile IA trebuie să beneficieze cât mai mulți oameni cu putință.*
- §15 Prosperitate partajată: *Prosperitatea economică creată de IA trebuie larg partajată, spre beneficiul întregii omeniri.*
- §16 Control uman: *Oamenii trebuie să aleagă cum și dacă își deleagă deciziile sistemelor IA, pentru îndeplinirea propriilor lor obiective.*

- §17 **Anti-subminare:** Puterea conferită prin controlul unor sisteme IA avansate trebuie să respecte și să amelioreze, nu să submineze, procesele sociale și civice de care depinde sănătatea societății.
- §18 **Cursa înarmărilor IA:** Trebuie evitată o cursă a înarmărilor cu arme autonome letale.

PROBLEME PE TERMEN LUNG

- §19 **Avertisment privind capacitățile:** Neexistând un consens în această privință, trebuie evitate presupunerile tranșante legate de limitele superioare ale capacităților viitoarelor IA.
- §20 **Importanță:** IA avansată poate reprezenta o profundă schimbare în istoria vieții pe Pământ, și trebuie planificată și gestionată cu atenția și resursele corespunzătoare.
- §21 **Riscuri:** Riscurile implicate de sistemele IA, în special cele care pot declanșa catastrofe sau amenința viața, trebuie supuse unor eforturi de prevedere și diminuare pe măsura impactului lor potențial.
- §22 **Autoperfecționare recursivă:** Sistemele IA proiectate pentru a se autoperfecționa sau autoreplica recursiv, cu riscul de a deveni rapid foarte numeroase sau foarte avansate, trebuie supuse unor măsuri stricte de siguranță și control.
- §23 **Binele comun:** Superintelența trebuie dezvoltată pentru a se afla numai în slujba unor idealuri etice larg împărtășite și în beneficiul întregii omeniri, nu doar al unui stat sau al unei organizații.

Lista de semnături a crescut spectaculos după ce am postat principiile online, iar acum conține peste o mie de cercetători IA și alte personalități de frunte. Dacă vrei și tu să semnezi, o poți face aici: <http://futureoflife.org/ai-principles>.

Am fost impresionați nu doar de nivelul ridicat al consensului asupra principiilor, ci și de forța acestora. Desigur, la prima vedere unele par de genul „Pacea, dragostea și maternitatea sunt lucruri de preț”, dar multe dintre ele își dovedesc eficacitatea atunci când formulezi negații ale lor. De pildă, „Superintelența e imposibilă” încalcă §19, iar „Este pierdere de vreme cercetarea pentru a reduce riscul la care IA expune existența” încalcă §21.

După cum puteți constata privind pe YouTube discuția despre perspectivele pe termen lung,¹¹ Elon Musk, Stuart Russell, Ray

Kurzweil, Demis Hassabis, Sam Harris, Nick Bostrom, David Chalmers, Bart Selman și Jaan Tallinn au fost cu toții de acord că superinteligenta va fi probabil dezvoltată, iar cercetările legate de siguranță sunt importante.

Sper ca principiile IA de la Asilomar să servească drept punct de pornire pentru discuții mai detaliate, care să conducă în cele din urmă la strategii și politici IA raționale. În acest spirit, Ariel Conn, responsabil cu relațiile publice la FLI, împreună cu Tucker Davey și alți membri ai echipei, au luat interviuri celor mai importanți cercetători IA în legătură cu principiile și felul în care le interpretează aceștia, în vreme ce David Stanley și echipa lui internațională de voluntari au tradus principiile în limbi de mare circulație.

OPTIMISM PRECAUT

După cum am mărturisit la începutul epilogului, sunt atât de optimist în privința viitorului vieții cum n-am mai fost de mult. V-am împărtășit povestea mea pentru a vă explica de ce.

Experiențele mele din ultimii câțiva ani mi-au sporit optimismul din două motive. În primul rând, am fost martor la o remarcabilă reunire a comunității IA pentru a accepta în mod constructiv provocările ce ne stau în față, adesea în colaborare cu gânditori din alte domenii. După întâlnirea de la Asilomar, Elon mi-a spus că îl uimea faptul că, în doar câțiva ani, dintr-o temă periferică, siguranța IA a devenit una centrală, iar eu însumi sunt la fel de uimit. Nu doar problemele care se pun pe termen scurt, prezentate în capitolul 3, sunt acum subiecte de discuție respectabile, dar și superinteligenta sau riscul la care e expusă existența, ca în principiile IA de la Asilomar. Acele principii nu puteau fi adoptate în Puerto Rico, cu doi ani mai devreme – cuvântul din scrisoarea deschisă care suna cel mai înspăimântător a fost „capcane”.

Îmi place să-i privesc pe oameni, iar în dimineața ultimei zile a conferinței de la Asilomar am stat în marginea sălii și m-am uitat



Figura 9.4 O comunitate tot mai numeroasă caută răspunsuri la Asilomar.

la participanții care urmăreau o discuție despre IA și legi. M-a cuprins o senzație de căldură difuză și am fost mișcat. Era cu totul altfel decât în Puerto Rico! Pe atunci, țin minte că-i priveam pe cei mai mulți din comunitatea IA cu un amestec de respect și teamă – nu ca pe o echipă adversă, ci ca pe un grup pe care eu și colegii mei preocupați de IA simțeam că trebuia să-l convingem. Acum era însă evident că făceam cu toții parte din *aceeași* echipă. După cum ați înțeles probabil citind această carte, tot nu am un răspuns la întrebarea cum ne putem făuri un viitor măreț cu IA, așa încât găsesc că e minunat să faci parte dintr-o comunitate în continuă creștere care caută răspunsul.

Al doilea motiv pentru care sunt mai optimist e acela că experiența FLI a fost mobilizatoare. Ceea ce făcuse să-mi dea lacrimile la Londra a fost o senzație de inevitabil: ne-ar putea aștepta un viitor îngrijorător și n-avem cum să ne împotrivim. Următorii trei ani mi-au risipit însă temerea fatalistă. Dacă o mână de voluntari neplătiți pot influența dezbaterea pe care o consider cea mai importantă a epocii noastre, imaginați-vă ce am putea face cu toții, dacă am lucra împreună!

Erik Brynjolfsson vorbea în cuvântarea sa de la Asilomar despre două tipuri de optimism. Există un optimism necondiționat, ca

în speranța că soarele va răsări și mâine, dar există și ceea ce el numea „optimism precaut” – speranța că se vor petrece lucruri bune dacă le plănuiești cu atenție și te străduiești pentru împlinirea lor. Acesta e optimismul pe care îl simt acum în privința viitorului vieții.

Așadar, ce poți face *tu* pentru a influența în mod pozitiv viitorul vieții, acum când intrăm în epoca inteligenței artificiale? Din motive pe care le voi explica imediat, cred că primul pas este acela de a deveni un optimist precaut, dacă nu cumva ești deja. Pentru a fi un bun optimist precaut, esențial este să ajungi la o perspectivă pozitivă asupra viitorului. Când studenții de la MIT vin să-mi ceară sfatul în privința carierei de urmat, încep de regulă prin a-i întreba unde se văd pe ei înșiși peste un deceniu. Dacă un student îmi răspunde „Am să fiu pesemne la spital, într-un salon de canceroși, sau în cimitir, după ce m-a călcat un autobuz”, nu-i acord mari șanse. Să-ți privești viitorul doar în culori întunecate e o abordare catastrofală pentru planificarea carierei! Să-ți dedici 100% din energie ca să eviți bolile și accidente este calea către ipohondrie și paranoia, nu către fericire. Aș vrea să-l aud pe student vorbind cu entuziasm despre obiectivele lui, după care putem discuta strategiile de a ajunge acolo, evitând capcanele.

Erik a subliniat că, în conformitate cu teoria jocurilor, perspectivele pozitive stau la baza unei mari părți a colaborărilor în lume, de la căsătorii și fuziuni de corporații până la hotărârile statelor independente de a forma Statele Unite ale Americii. În fond, de ce să sacrifici ceva ce ai, dacă nu-ți poți închipui un câștig mai mare pe care l-ai obține? Prin urmare, ar trebui să ne imaginăm un viitor pozitiv nu numai pentru noi înșine, dar și pentru societate și omenire. Cu alte cuvinte, avem nevoie de mai multă speranță existențială. După cum îmi reamintește adesea Meia, de la Frankenstein la Terminator, viziunile despre viitor în literatură și film sunt predominant distopice. Altfel spus, noi, ca societate, ne plănuim viitorul cam la fel de rău ca ipoteticul meu student de la MIT. De aceea avem nevoie de mai mulți optimiști precauți. Și tot de aceea v-am îndemnat în cartea de față să vă

gândiți la viitorul pe care vi-l *doriți*, nu la viitorul de care vă temeți, așa încât să găsim obiective comune pe care să le plănuim și cărora să le dedicăm eforturile noastre.

Am văzut în această carte că IA va aduce probabil cu sine posibilități mărețe și provocări extreme. O strategie care ne va ajuta pesemne în practic toate provocările legate de IA este aceea de a acționa împreună pentru a ameliora societatea umană *înainte* ca IA să ia avânt. Ar fi bine să-i educăm pe tineri pentru a face ca tehnologia să fie robustă și benefică înainte să i se cedeze acesteia o mare parte din putere. Ar fi bine să ne modernizăm legile înainte ca tehnologia să le facă să devină desuete. Ar fi bine să rezolvăm conflictele internaționale înainte de escaladarea unei curse a înarmărilor cu armament autonom. Ar fi bine să făurim o economie care asigură tuturor prosperitate înainte ca IA să poată amplifica inegalitățile. Am trăi mai bine într-o societate în care rezultatele cercetărilor privind siguranța IA sunt puse în aplicare, nu ignorate. Și, privind spre viitor la provocările legate de IAG supraumană, ar fi bine să cădem de acord măcar asupra câtorva standarde etice de bază, înainte de a pune mașini puternice să învețe aceste standarde. Într-o lume polarizată și haotică, oamenii care au puterea de a folosi IA în scopuri malefice vor fi tot mai motivați și mai capabili s-o facă, iar echipele aflate în competiție pentru construirea IAG vor fi supuse unei mari presiuni ca, în loc să coopereze în problemele de siguranță, să trișeze. În concluzie, dacă putem edifica o societate umană mai armonioasă caracterizată prin cooperare pentru atingerea unor obiective comune, sporesc șansele ca revoluția IA să urmeze un curs favorabil.

Cu alte cuvinte, una dintre cele mai bune căi de a ameliora viitorul vieții este aceea de a ameliora ziua de mâine. Ai posibilitatea s-o faci în multe moduri. Poți, desigur, vota la urne, spunând politicienilor ce crezi despre educație, confidențialitate, armele autonome letale, șomaj tehnologic și altele. Dar votezi și zi de zi prin ceea ce alegi să cumperi, prin știrile pe care alegi să le urmărești, prin ce alegi să împărtășești celorlalți, prin exemplul pe care alegi să-l dai. Vrei să întrerupi conversațiile celorlalți controlân-

du-le telefoanele inteligente, sau vrei ca tehnologia să fie folosită într-un mod planificat și deliberat? Vrei ca tu să posezi tehnologia, sau vrei ca tehnologia să te posede pe tine? Ce ai vrea să însemne a fi om în epoca IA? Te invit să discuți despre toate acestea cu cei din jurul tău – conversația nu e doar importantă, dar și fascinantă.

Pe măsură ce modelăm epoca inteligenței artificiale, suntem răspunzători pentru viitorul vieții. Deși la Londra am plâns, simt acum că nimic nu e inevitabil în privința acestui viitor, și știu că e mult mai ușor să-l influențezi decât îmi închipuiam. Viitorul nostru nu e săpat în piatră, urmând să ni se întâmple negreșit – noi suntem cei care îl construim. Să construim împreună unul rodnic!

Note

CAPITOLUL 1

1. „The AI Revolution: Our Immortality or Extinction?“, *Wait But Why* (27 ianuarie 2015), la <http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html>.
2. Această scrisoare deschisă, „Research Priorities for Robust and Beneficial Artificial Intelligence“, poate fi găsită la <http://futureoflife.org/ai-open-letter/>.
3. Exemplu de alarmism clasic privind roboții în mediile de informare: Ellie Zolfagharifard, „Artificial Intelligence «Could Be the Worst Thing to Happen to Humanity»“, *Daily Mail*, 2 mai 2014, <http://tinyurl.com/hawkingbots>.

CAPITOLUL 2

1. Observații despre originea termenului IAG: <http://wp.goertzel.org/who-coined-the-term-ai>.
2. Hans Moravec, „When Will Computer Hardware Match the Human Brain“, *Journal of Evolution and Technology* (1998), vol. 1.
3. În figura care prezintă evoluția puterii de calcul, datele dinainte de 2011 sunt luate din cartea lui Ray Kurzweil *How to Create a Mind*, iar datele ulterioare sunt calculate din referințele de la <https://en.wikipedia.org/wiki/FLOPS>.
4. David Deutsch prezintă calculul cuantic drept o dovadă a universurilor paralele în cartea sa *The Fabric of Reality* (Allen Lane, Londra, 1997). Opinia mea, după care universurile paralele cuantice reprezintă al treilea din cele patru niveluri ale multiversului, se găsește în precedenta mea carte, *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality* (Knopf, New York, 2014).

5. Pentru succesul echipei Google în descrierea imaginilor cu ajutorul IA vezi <https://arxiv.org/pdf/1411.4555.pdf>.

CAPITOLUL 3

1. Priviți „Google DeepMind's Deep Q-learning Playing Atari Breakout“ pe YouTube la adresa: <https://tinyurl.com/atari1ai>.
2. Vezi Volodymyr Mnih *et al.*, „Human-Level Control Through Deep Reinforcement Learning“, *Nature* 518 (26 februarie 2015), disponibil online la <http://tinyurl.com/atari1paper>.
3. Document video cu robotul Big Dog în acțiune: <https://www.youtube.com/watch?v=W1czBcnXrWw>.
4. Pentru reacții la mutarea senzațională a lui AlphaGO, vezi „Move 37!! Lee Sedol vs AlphaGo Match 2“, la: <https://www.youtube.com/watch?v=JN-rXgpSEEIE>.
5. Demis Hassabis prezentând reacțiile la AlphaGo ale unor jucători de go umani: <https://www.youtube.com/watch?v=otJKzpNWZT4>.
6. Pentru progresele recente în traducerea cu ajutorul mașinilor, vezi Gideon Lewis-Kraus, „The Great AI Awakening“, *New York Times Magazine* (14 decembrie 2016), disponibil online la <http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html>. Google Translate e disponibil la <https://translate.google.com>.
7. Competiția Provocarea Schemei Winograd: <http://tinyurl.com/winogradchallenge>.
8. Document video cu explozia rachetei Ariane 5: <https://www.youtube.com/watch?v=qnHn8W1Em6E>.
9. Raportul privind explozia rachetei Ariane 5: <http://tinyurl.com/arianeflop>.
10. Raportul NASA privind eșecul sondei Mars Climate Orbiter: <http://tinyurl.com/marsflop>.
11. Cea mai amănunțită și consistentă relatare a cauzelor eșecului misiunii Mariner 1: <http://tinyurl.com/marinerflop>.
12. O descriere amănunțită a eșecului misiunii sovietice Phobos 1 poate fi găsită în Wesley T. Huntress Jr. și Mikhail Ya. Marov, *Soviet Robots in the Solar System* (Praxis Publishing, New York, 2011), p. 308.
13. Despre felul în care un software neverificat a făcut ca Knight Capital să piardă 440 de milioane de dolari în 45 de minute: <http://tinyurl.com/knightflop1> și <http://tinyurl.com/knightflop2>.
14. Raportul guvernamental privind „crahul fulger“: „Findings Regarding the Market Events of May 6, 2010“ (30 septembrie 2010), la <http://tinyurl.com/flashcrashreport>.

15. Exemple de folosire a imprimantelor 3-D pentru construcții (<https://www.youtube.com/watch?v=SOBzNdyRTBs>), pentru dispozitive micromecanice (<http://tinyurl.com/tinyprinter>) și pentru multe alte lucruri intermediare (<https://www.youtube.com/watch?v=xVU4FLrsPXs>).
16. Harta lumii pe care sunt trecute atelierele-laborator: <https://www.fablabs.io/map>.
17. Articol despre uciderea lui Robert Williams de către un robot industrial: <http://tinyurl.com/williamsaccident>.
18. Articol despre uciderea lui Kenji Urada de către un robot industrial: <http://tinyurl.com/uradaaccident>.
19. Articol despre uciderea unui lucrător de la Volkswagen de către un robot industrial: <http://tinyurl.com/baunatalaccident>.
20. Raport al guvernului SUA privind accidentele de muncă: https://www.osha.gov/dep/fatcat/dep_fatcat.html.
21. Statistica accidentelor rutiere: <http://tinyurl.com/roadsafety2> și <http://tinyurl.com/roadsafety3>.
22. Despre primul accident al unei mașini Tesla autopilotate, vezi Andrew Buncombe, „Tesla Crash: Driver Who Died While on Autopilot Mode «Was Watching Harry Potter»“, *Independent* (1 iulie 2016), <http://tinyurl.com/teslacrashstory>. Raportul Administrației Naționale pentru Traficul Rutier: <http://tinyurl.com/teslacrashreport>.
23. Amănunte despre catastrofa feribotului Herald of Free Enterprise găsiți în R.B. Whittingham, *The Blame Machine: Why Human Error Causes Accidents* (Elsevier, Oxford, 2004).
24. Documentar despre prăbușirea zborului Air France 447: <https://www.youtube.com/watch?v=dpPkp8OGQFI>; raport despre accident: <http://tinyurl.com/af447report>; analiză externă: <http://tinyurl.com/thomsonarticle>.
25. Raport oficial despre pana de curent din Statele Unite și Canada: <http://tinyurl.com/uscanadablackout>.
26. Raport final al comisiei prezidențiale despre accidentul de la Three Mile Island: <http://www.threemileisland.org/downloads/188.pdf>.
27. Studiu olandez care arată că IA poate rivaliza cu radiologii umani în diagnosticarea cancerului de prostată pe baza IRM: <http://tinyurl.com/prostate-ai>.
28. Studiu Stanford care arată că IA îi poate depăși pe anatomopatologii umani în diagnosticarea cancerului la plămân: <http://tinyurl.com/lungcancer-ai>.
29. Investigație privind accidentele provocate de Therac-25: <http://tinyurl.com/theracfailure>.
30. Raport privind supradozele letale de la Institutul Oncologic Național din Panama: <http://tinyurl.com/cobalt60accident>.
31. Studiu privind accidentele provocate de chirurgia robotizată: <https://arxiv.org/abs/1507.03518>.

32. Articol despre numărul deceselor provocate de malpraxisul medical în spitale: <http://tinyurl.com/medaccidents>.
33. Yahoo a luat măsuri de siguranță după ce în 2013 a anunțat că hackerii au pătruns în peste un miliard (!) de conturi: <https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html>.
34. Articol din *New York Times* despre achitarea, iar apoi condamnarea asasinului care făcea parte din Ku Klux Klan: <http://tinyurl.com/kkkacquittal>.
35. Studiul efectuat de Danziger *et al.* în 2011 (<http://www.pnas.org/content/108/17/6889.full>), susținând că judecătorii cărora le e foame sunt mai aspri, a fost criticat de Keren Weinshall-Margela și John Shapard (<http://www.pnas.org/content/108/42/E833.full>), dar Danziger *et al.* insistă că afirmațiile lor rămân valabile (<http://www.pnas.org/content/108/42/E834.full>).
36. *Pro Publica* afirmă că software-ul folosit pentru a prezice recidivismul nu e imparțial: <http://tinyurl.com/robojudge>.
37. Folosirea IRMf și a altor tehnici de scanare cerebrală ca probă în procese este extrem de controversată, la fel și încrederea în asemenea tehnici, deși multe echipe susțin că acuratețea e de peste 90%: <http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00709/full>.
38. Rețeaua de televiziune PBS a făcut filmul *The Man Who Saved the World* despre incidentul în care Vasili Arhipov a împiedicat de unul singur o lovitură nucleară sovietică: <https://www.youtube.com/watch?v=4VPY2SgyG5w>.
39. Povestea lui Stanislav Petrov, care a considerat falsă avertizarea că urmează un atac nuclear american, a fost ecranizată în filmul *The Man Who Saved the World* (a nu se confunda cu filmul cu același titlu din nota precedentă), iar ONU i-a acordat lui Petrov distincția „Cetățeanul Lumii”: <https://www.youtube.com/watch?v=IncSjwWQHMo>.
40. Scrisoare deschisă a cercetătorilor din IA și robotică privind armele autonome: <http://futureoflife.org/open-letter-autonomous-weapons/>.
41. Un oficial american care pare să-și dorească o cursă a înarmărilor IA: <http://tinyurl.com/workquote>.
42. Studiu despre inegalitatea averilor în Statele Unite începând din 1913: <http://gabriel-zucman.eu/files/SaezZucman2015.pdf>.
43. Organizația Oxfam despre inegalitatea averilor: <http://tinyurl.com/oxfam2017>.
44. Pentru o prezentare a ipotezei că inegalitatea e indusă de tehnologie, vezi Erik Brynjolfsson și Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (Norton, New York, 2014).
45. Articol din *The Atlantic* despre scăderea salariilor celor mai puțin instruiți: <http://tinyurl.com/wagedrop>.
46. Datele din grafic sunt luate din Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, Emmanuel Saez și Gabriel Zucman, *The World Wealth and*

- Income Database* (<http://www.wid.world>, 31/10/2016), incluzând câștigurile de capital.
47. Prezentare a lui James Manyika: http://futureoflife.org/data/PDF/james_manyika.pdf.
 48. Previziuni privind viitorul automatizării joburilor elaborate de Universitatea Oxford (<http://tinyurl.com/automationoxford>) și de McKinsey (<http://tinyurl.com/automationmckinsey>).
 49. Document video cu un bucătar robot: <https://www.youtube.com/watch?v=fE6i2OO6Y6s>.
 50. Marin Solačić a examinat aceste opțiuni la seminarul din 2016 „Computers Gone Wild: Impact and Implications of Developments in Artificial Intelligence on Society”: <http://futureoflife.org/2016/05/06/computers-gone-wild/>.
 51. Propunerile lui Andrew McAfee pentru crearea mai multor joburi: http://futureoflife.org/data/PDF/andrew_mcafee.pdf.
 52. În plus față de numeroasele articole academice susținând că șomajul tehnologic „de data asta e altfel”, documentul video „Humans need not apply” prezintă pe scurt aceeași idee: <https://www.youtube.com/watch?v=7Pq-S557XQU>.
 53. Oficiul Statelor Unite pentru Statistica Muncii: <http://www.bls.gov/cps/cpsaat11.htm>.
 54. Argument că „de data asta e altfel” pentru șomajul tehnologic: Federico Pistono, *Robots Will Steal Your Job, but That's OK* (2012), <http://robotswillstealyourjob.com>.
 55. Evoluția șeptelului cabalin american: <http://tinyurl.com/horsedecline>.
 56. Meta-analiză arătând cum afectează șomajul starea de bine: Maike Luhmann et al., „Subjective Well-Being and Adaptation to Life Events: A Meta-Analysis”, *Journal of personality and social psychology* 102, nr. 3 (2012), p. 592, disponibil online la <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289759>.
 57. Studii privind sursele senzației de bunăstare: Angela Duckworth, Tracy Steen și Martin Seligman, „Positive Psychology in Clinical Practice”, *Annual Review of Clinical Psychology* 1 (2005), pp. 629-651, online la <http://tinyurl.com/wellbeingduckworth>; Weiting Ng și Ed Diener, „What Matters to the Rich and the Poor? Subjective Well-Being, Financial Satisfaction, and Postmaterialist Needs Across the World”, *Journal of personality and social psychology* 107, nr. 2 (2014), p. 326, online la <http://psycnet.apa.org/journals/psp/107/2/326>; Kirsten Weir, „More than Job Satisfaction”, *Monitor on Psychology* 44, nr. 11 (decembrie 2013), online la <http://www.apa.org/monitor/2013/12/job-satisfaction.aspx>.
 58. Înmulțind circa 10^{22} neuroni cu circa 10^4 conexiuni pe neuron și cu circa 0 (10^0) activare pe neuron în fiecare secundă, ar rezulta că aproximativ

- 10¹⁵ FLOPS (1 petaFLOPS) sunt de ajuns pentru a simula un creier uman, dar rămân nelămurite încă multe lucruri, între care ritmul activărilor și întrebarea dacă nu cumva părțile mici ale neuronilor și sinapselor trebuie și ele simulate. Informaticianul Dharmendra Modha de la IBM a estimat că e nevoie de 38 de petaFLOPS (<http://tinyurl.com/javl43>), în vreme ce specialistul în neuroștiințe Henry Markram a estimat că e nevoie de circa 1 000 de petaFLOPS (<http://tinyurl.com/6rpohqv>). Cercetătorii IA Katja Grace și Paul Christiano au susținut că aspectul cel mai costisitor al simulării nu e calculul, ci comunicarea, iar și această sarcină poate fi în linii mari îndeplinită de cele mai bune supercalculatoare actuale: <http://aiimpacts.org/about>.
59. Pentru o estimare interesantă a puterii de calcul a creierului uman: Hans Moravec, „When Will Computer Hardware Match the Human Brain”, *Journal of Evolution and Technology*, vol. 1 (1998).

CAPITOLUL 4

1. Pentru un document video cu prima pasăre mecanică, vezi Markus Fischer, „A Robot That Flies Like a Bird”, Conferință TED, iulie 2011, la https://www.ted.com/talks/a_robot_that_flies_like_a_bird.

CAPITOLUL 5

1. Ray Kurzweil, *The Singularity Is Near* (Viking Press, New York, 2005).
2. Scenariul lui Ben Goertzel „Nanny AI” e prezentat la https://wiki.lesswrong.com/wiki/Nanny_AI.
3. Pentru o discuție despre relația dintre mașini și oameni, vezi Benjamin Wallace-Wells, „Boyhood”, *New York Magazine* (20 mai 2015), online la <http://tinyurl.com/aislaves>.
4. Crima împotriva minții e discutată în cartea lui Nick Bostrom *Superintelligence*, și mai detaliat în acest articol recent: Nick Bostrom, Allan Dafoe și Carrick Flynn, „Policy Desiderata in the Development of Machine Superintelligence” (2016), <http://www.nickbostrom.com/papers/aipolicy.pdf>.
5. Matthew Schofield, „Memories of Stasi Color German's View of U.S. Surveillance Programs”, *McClatchy DC Bureau* (26 iunie 2013), online la <http://www.mcclatchydc.com/news/nation-world/national/article24750439.html>.
6. Pentru reflecții pătrunzătoare privind felul în care oamenii pot fi îndemnați să acționeze pentru a se ajunge la rezultate pe care nimeni nu și le dorește, recomand „Meditations on Molloch”, <http://slatestarcodex.com/2014/07/30/meditations-on-moloch>.
7. Pentru o listă a cazurilor în care războiul nuclear a fost cât pe ce să fie declanșat accidental, vezi Institutul pentru Viitorul Vieții, „Accidental Nuclear War: A Timeline of Close Calls”, online la <http://tinyurl.com/nukeoops>.

8. Pentru despăgubiri acordate victimelor testelor nucleare americane, vezi site-ul Departamentului de Justiție al Statelor Unite, „Awards to Date 4/24/2015”, la <https://www.justice.gov/civil/awards-date-04242015>.
9. *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack*, aprilie 2008, disponibil online la http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf.
10. Cercetări independente ale oamenilor de știință americani și sovietici i-au prevenit pe Reagan și Gorbaciov în privința riscului iernii nucleare: P.J. Crutzen și J.W. Birks, „The Atmosphere After a Nuclear War: Twilight at Noon”, *Ambio* 11, nr. 2/3 (1982), pp. 114–125; R.P. Turco, O.B. Toon, T.P. Ackerman, J.B. Pollack și C. Sagan, „Nuclear Winter: Global Consequences of Multiple Nuclear Explosions”, *Science* 222 (1983), pp. 1283–1292; V.V. Aleksandrov și G.L. Stenchikov, „On the Modeling of the Climatic Consequences of the Nuclear War”, *Proceeding on Applied Mathematics* (Moscova, Centrul de Calcul al Academiei de Științe a URSS) 21; A. Robock, „Snow and Ice Feedbacks Prolong Effects of Nuclear Winter”, *Nature* 310 (1984), pp. 667–670.
11. Calcul al efectelor războiului nuclear global asupra climei: A. Robock, L. Oman și L. Stenchikov, „Nuclear Winter Revisited with a Modern Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences”, *Journal of Geophysical Research* 12 (2007), p. D13107.

CAPITOLUL 6

1. Pentru informații suplimentare, vezi Anders Sandberg, „Dyson Sphere FAQ”, la <http://tinyurl.com/dysonsph>.
2. Articolul lui Freeman Dyson despre sferele care-i poartă numele: Freeman Dyson, „Search for Artificial Stellar Sources of Infrared Radiation”, *Science*, vol. 131 (1959), pp. 1667–1668.
3. Louis Crane și Shawn Westmoreland explică motorul cu găuri negre propus de ei în „Are Black Hole Starship Possible?”, la <http://arxiv.org/pdf/0908.1803.pdf>.
4. Pentru un grafic produs de CERN care trece în revistă particulele elementare cunoscute, vezi <http://tinyurl.com/cernparticles>.
5. Acest document video cu un prototip Orion nenuclear ilustrează ideea de propulsie cu bombe nucleare: <https://www.youtube.com/watch?v=E-3Lxx2VAYi8>.
6. O prezentare accesibilă a navigației pe bază de laser: Robert L. Forward, „Roundtrip Interstellar Travel Using Laser-Pushed Lightsails”, *Journal of Spacecrafts and Rockets* 21, nr. 2 (martie-aprilie 2014), disponibil online la <http://www.lunarsail.com/LightSail/rit-1.pdf>.

7. Jay Olson analizează expansiunea cosmică a civilizațiilor în „Homogenous Cosmology with Aggressively Expanding Civilizations”, *Classical and Quantum Gravity* 32 (2015), disponibil online la <http://arxiv.org/abs/1411.4359>
8. Prima analiză științifică a viitorului nostru îndepărtat: Freeman J. Dyson, „Time Without End: Physics and Biology in an Open Universe”, *Reviews of Modern Physics* 51, nr. 3 (1979), p. 447, disponibil online la http://blog.regehr.org/extra_files/dyson.pdf.
9. Formula lui Seth Lloyd menționată mai sus ne spune că efectuarea unei operații de calcul în intervalul de timp τ consumă o energie $E \geq h/4\tau$, unde h este constanta lui Planck. Dacă vrem să efectuăm N operații una după alta (în serie) în timpul T , atunci $\tau = T/N$, deci $E/N \geq hN/4T$, de unde rezultă că putem efectua $N \leq 2ET/h$ operații în serie folosind energia E în timpul T . Dacă împărțim energia între n calcule paralele, ele pot fi efectuate mai lent și mai eficient, așa încât $N \leq 2\sqrt{ETn}/h$. Nick Bostrom estimează că simularea unei vieți umane de 100 de ani necesită aproximativ $N = 10^{27}$ operații.
10. Un raționament care arată că originea vieții presupune un accident extrem de rar, plasând vecinii noștri cei mai apropiați la o distanță de peste 10^{1000} metri, poate fi găsit în acest document video al fizicianului și astrobiologului de la Princeton Edwin Turner: „Improbable Life: An Unappealing but Plausible Scenario for Life's Origin on Earth”, la <https://www.youtube.com/watch?v=Bt6n6Turbeg>.
11. Eseu al lui Martin Rees despre căutarea inteligenței extraterestre: <https://www.edge.org/annual-question/2016/response/26665>.

CAPITOLUL 7

1. O prezentare accesibilă a cercetărilor lui Jeremy England privind „adaptarea determinată de disipare” poate fi găsită în Natalie Wolchover, „A New Physics Theory of Life”, *Scientific American* (28 ianuarie 2014), disponibil online la <https://www.scientificamerican.com/article/a-new-physics-theory-of-life/>. Cartea lui Ilya Prigogine și Isabelle Stengers *Order Out of Chaos: Man's New Dialogue with Nature* (Bantam, New York, 1984) stabilește în mare măsură fundamentele acestor cercetări.
2. Pentru bazele fiziologice ale sentimentelor, vezi: William James, *Principles of Psychology* (New York, Henry Holt & Co., 1890); Robert Ornstein, *Evolution of Consciousness: The Origins of the Way We Think* (New York, Simon & Schuster, 1992); António Damásio, *Descartes' Error: Emotion, Reason, and the Human Brain* (Penguin, New York, 2005) [ed. rom. *Eroarea lui Descartes. Emoțiile, rațiune și creierul uman*, Humanitas, București, 2005]; António Damásio, *Self Comes to Mind: Constructing the Conscious Brain* (Vintage, New York, 2012) [ed. rom. *Sinele. Construirea creierului inteligent*, Humanitas, București, 2019].
3. Eliezer Yudkowsky a analizat alinierea obiectivelor unei IA prietenoase

- nu cu obiectivele noastre actuale, ci cu *voința noastră extrapolată coerent* (*coherent extrapolated volition* – CEV). În linii mari, aceasta e definită drept ceea ce ar vrea o versiune idealizată a noastră dacă ar ști mai multe și ar gândi mai repede, reprezentând mai curând ceea ce am vrea noi să fim. Yudkowsky a început să critice CEV la scurt timp după ce a publicat articolul despre ea (<http://intelligence.org/files/CEV.pdf>), pe motiv că e greu de operat cu ea și nu e limpede dacă poate converge către ceva bine definit.
4. În abordarea învățării prin consolidare inversă, o idee esențială este că IA nu încearcă să maximizeze satisfacția îndeplinirii obiectivelor pentru sine, ci pentru proprietarul ei uman. Ea este deci îndemnată să fie precută atunci când nu știe clar ce vrea proprietarul, și să se străduiască să afle. Ar trebui de asemenea să nu aibă de obiectat dacă proprietarul o scoate din funcțiune, fiindcă asta înseamnă că i-a înțeles greșit dorința.
 5. Articolul lui Steve Omohundro despre apariția obiectivelor unei IA poate fi găsit la <http://tinyurl.com/omohundro2008>. El a fost publicat inițial în *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, ed. Pei Wang, Ben Goertzel și Stan Franklin (IOS, Amsterdam, 2008), pp. 483–492.
 6. O carte despre ce se întâmplă când inteligența e folosită pentru a se supune orbește ordinelor, fără să-și pună probleme privind baza lor etică: Hanna Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (Penguin, New York, 1963). O dilemă înrudită se leagă de propunerea lui Eric Drexler (<http://www.fhi.ox.ac.uk/reports/2015-3.pdf>) de a ține superinteligența sub control prin compartimentarea ei în părți simple, așa încât nici una dintre acestea să nu înțeleagă imaginea de ansamblu. Dacă metoda ar funcționa, s-ar obține un instrument incredibil de puternic fără vreo busolă morală intrinsecă, care ar pune în aplicare toate capriciile proprietarului ei, fără scrupule de ordin moral. Ar semăna cu o birocrație compartimentată dintr-o dictatură distopică: o parte construiește arme fără să știe că ele vor fi folosite, alta execută deșinuți fără să știe că au fost condamnați etc.
 7. O variantă modernă a Regulii de Aur este ideea lui John Rawls că o situație ipotetică e acceptabilă dacă nimeni n-ar schimba-o fără să știe dinainte cine se va confrunta cu ea.
 8. De pildă, coeficientul de inteligență al multor funcționari de rang înalt ai lui Hitler pare să fi fost destul de ridicat. Vezi „How Accurate Were the IQ Scores of the High-Ranking Third Reich Officials Tried at Nuremberg?”, *Quora*, disponibil la <http://tinyurl.com/nurembergiq>.

CAPITOLUL 8

1. Articolul despre conștiință al lui Stuart Sutherland e foarte amuzant: *Macmillan Dictionary of Psychology* (Macmillan, Londra, 1989).

2. Erwin Schrödinger, unul dintre părinții mecanicii cuantice, a făcut această remarcă în cartea sa *Mind and Matter* gândindu-se la trecut – și la ce s-ar fi întâmplat dacă viața conștientă n-ar fi evoluat. Pe de altă parte, ascensiunea IA face să apară în mod logic posibilitatea de a avea o piesă care să se joace în fața unei săli pustii în viitor.
3. *Stanford Encyclopedia of Philosophy* dă diferite definiții și întrebări ale cuvântului „conștiință”: <http://tinyurl.com/stanfordconsciousness>.
4. Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (Harper-Collins, New York), p. 116 [ed. rom. *Homo Deus. Scurtă istorie a viitorului*, Polirom, Iași, 2018].
5. O excelentă prezentare a Sistemului 1 și a Sistemului 2: Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus & Giroux, New York, 2011).
6. Vezi Christof Koch, *The Quest for Consciousness: A Neurobiological Approach* (W.H. Freeman, New York, 2004).
7. Suntem pesemne conștienți doar de o infimă parte (10–50 biți) din informația care intră în creierul nostru în fiecare secundă: K. Küpfmüller, „Nachrichtenverarbeitung im Menschen”, în *Taschenbuch der Nachrichtenverarbeitung*, ed. K. Steinbuch, (Springer-Verlag, Berlin, 1962), pp. 1481–1502; T. Nørretranders, *The User Illusion: Cutting Consciousness Down to Size* (Viking, New York, 1991).
8. Michio Kaku, *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind* (Doubleday, New York, 2014); Jeff Hawkins și Sandra Blakeslee, *On Intelligence* (Times Books, New York, 2007); Stanislas Dehaene, Michel Kerszberg și Jean-Pierre Changeux, „A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks”, *Proceedings of the National Academy of Sciences* 95 (1998), pp. 14529–14534.
9. Document video despre celebrul experiment al lui Penfield „simt miros de pâine prăjită”: <https://www.youtube.com/watch?v=mSN86kphL68>. Detalii despre cortexul somatosenzorial: Elaine Marieb și Katja Hoehn, *Anatomy & Physiology*, ed. a III-a (Pearson, Upper Saddle River, NJ, 2008), pp. 391–395.
10. Studiul corelativelor neuronale ale conștiinței a intrat în ultimii ani în centrul atenției – vezi, de pildă, Geraint Rees, Gabriel Kreiman și Christof Koch, „Neural Correlates of Consciousness in Humans”, *Nature Reviews Neuroscience* 3 (2002), pp. 261–270, și Thomas Metzinger, *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (MIT Press, Cambridge, MA, 2000).
11. Cum funcționează suprimarea continuă a instantaneelor: Christof Koch, *The Quest for Consciousness: A Neurobiological Approach* (W.H. Freeman, New York, 2004); Christof Koch și Naotsugu Tsuchiya, „Continuous Flash

- Suppression Reduces Negative Afterimages", *Nature Neuroscience* 8 (2005), pp. 1096–1101.
12. Christof Koch, Marcello Massimini, Melanie Boly și Giulio Tononi, „Neural Correlates of Consciousness: Progress and Problems”, *Nature Reviews Neuroscience* 17 (2016), p. 307.
 13. Vezi Koch, *The Quest for Consciousness*, p. 260 și discuția din *Stanford Encyclopedia of Philosophy*, <http://tinyurl.com/consciousnessdelay>.
 14. Despre sincronizarea percepției conștiente: David Eagleman, *The Brain: The Story of You* (Pantheon, New York, 2015) [ed. rom. Creierul. Povestea noastră, Humanitas, București, 2018] și *Stanford Encyclopedia of Philosophy*, <http://tinyurl.com/consciousnesssync>.
 15. Benjamin Libet, *Mind Time: The Temporal Factor in Consciousness* (Harvard University Press, Cambridge, MA, 2004); Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze și John-Dylan Haynes, „Unconscious Determinants of Free Decisions in the Human Brain”, *Nature Neuroscience* 11 (2008), pp. 543–545, online la <http://www.nature.com/neuro/journal/v11/n5/full/nn.2112.html>.
 16. Exemple de abordări teoretice recente ale conștiinței: Daniel Dennett, *Consciousness Explained* (Back Bay Books, 1992); Bernard Baars, *In the Theater of Consciousness: The Workspace of the Mind* (Oxford University Press, New York, 2001); Christof Koch, *The Quest for Consciousness: A Neurobiological Approach* (W.H. Freeman, New York, 2004); Gerald Edelman și Giulio Tononi, *A Universe Of Consciousness How Matter Becomes Imagination: How Matter Becomes Imagination* (Hachette, New York, 2008); António Damásio, *Self Comes to Mind: Constructing the Conscious Brain* (Vintage, New York, 2012); Stanislas Dehaene, *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (Viking, New York, 2014); Stanislas Dehaene, Michel Kerszberg și Jean-Pierre Changeux, „A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks”, *Proceedings of the National Academy of Sciences* 95 (1998), pp. 14529–14534; Stanislas Dehaene, Lucie Charles, Jean-Rémi King și Sébastien Marti, „Toward a Computational Theory of Conscious Processing”, *Current Opinion in Neurobiology* 25 (2014), pp. 760–784.
 17. Pentru o discuție aprofundată privind diferitele întrebări ale termenului „emergență” în fizică și filozofie: David Chalmers, <http://cse3521.artifice.cc/Chalmers-Emergence.pdf>.
 18. Argumentul meu conform căruia conștiința este ceea ce simte informația atunci când e procesată în anumite moduri complexe: <https://arxiv.org/abs/physics/0510188>; <https://arxiv.org/abs/0704.0646>; Max Tegmark, *Our Mathematical Universe* (Knopf, New York, 2014). David Chalmers exprimă o idee asemănătoare în cartea sa din 1996 *The Conscious Mind*: „Experiența este informația privită din interior; fizica este informația privită din exterior.”

19. Adenauer Casali *et al.*, „A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior“, *Science Translational Medicine* 5 (2013), 198ra105, online la <http://tinyurl.com/zapzip>.
20. Teoria informației integrate nu se aplică sistemelor continue: <https://arxiv.org/abs/1401.1219>; <http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00063/full>; <https://arxiv.org/abs/1601.02626>.
21. Interviu cu Clive Wearing, a cărui memorie pe termen scurt durează doar circa 30 de secunde: <https://www.youtube.com/watch?v=WmzU47i2xgw>.
22. Critica lui Scott Aaronson la adresa IIT: <http://www.scottaaronson.com/blog/?p=1799>.
23. Critica lui Cerrullo la adresa IIT, susținând că integrarea nu e o condiție suficientă pentru conștiință: <http://tinyurl.com/cerrullocritique>.
24. Predicția IIT că oamenii simulați vor fi zombi: <http://rstb.royalsocietypublishing.org/content/370/1668/20140167>.
25. Critica lui Shanahan la adresa IIT: <http://arxiv.org/ftp/arxiv/papers/1504/1504.05696.pdf>.
26. Vedere oarbă: <http://tinyurl.com/blindsight-paper>.
27. Poate că suntem conștienți doar de o parte minusculă (10-50 biți) din informația care intră în creierul nostru în fiecare secundă: Küpfmüller, „Nachrichtenverarbeitung im Menschen“; Nørretranders, *The User Illusion*.
28. Argumente pro și contra „conștiinței fără acces“: Victor Lamme, „How Neuroscience Will Change Our View on Consciousness“, *Cognitive Neuroscience* (2010), pp. 204-220, online la <http://www.tandfonline.com/doi/abs/10.1080/17588921003731586>.
29. „Selective Attention Test“, <https://www.youtube.com/watch?v=vJG698U2Mvo>.
30. Vezi Lamme, „How Neuroscience Will Change Our View on Consciousness“.
31. Acesta și alte subiecte înrudite sunt discutate în cartea lui Daniel Dennett *Consciousness Explained*.
32. Vezi Kahneman, *Thinking, Fast and Slow*.
33. *Stanford Encyclopedia of Philosophy* despre liberul-arbitru: <https://plato.stanford.edu/entries/freewill>.
34. Document video cu Seth Lloyd explicând de ce o IA va simți că are liber-arbitru: <https://www.youtube.com/watch?v=Epj3DF8jDWk>.
35. Vezi Steven Weinberg, *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature* (Pantheon, New York, 1992) [ed. rom. *Visul unei teorii finale. În căutarea legilor ultime ale naturii*, Humanitas, București, 2010].
36. Prima analiză științifică a viitorului nostru îndepărtat: Freeman J. Dyson, „Time Without End: Physics and Biology in an Open Universe“, *Reviews of Modern Physics* 51, nr. 3 (1979), p. 447, disponibil online la <http://tinyurl.com/dysonfuture>.

EPILOG

1. Scrisoarea deschisă (<http://futureoflife.org/ai-open-letter/>) apărută în urma conferinței din Puerto Rico susținea că studierea modului de a face sistemele IA robuste și benefice este deopotrivă necesară și urgentă, și că există direcții concrete de cercetare care pot fi urmate încă de pe acum, așa cum e exemplificat în acest document despre prioritățile în cercetare: http://futureoflife.org/data/documents/research_priorities.pdf.
2. Interviu meu cu Elon Musk despre siguranța IA: <https://www.youtube.com/watch?v=rBwoeoZTY-g>
3. Compilație video cu aproape toate încercările de aterizare ale rachetei SpaceX, culminând cu prima amerizare reușită: <https://www.youtube.com/watch?v=AllaFzIPaG4>.
4. Elon Musk anunță pe tweeter bursele pentru siguranța IA: <https://twitter.com/elonmusk/status/555743387056226304>.
5. Elon Musk anunță pe tweeter scrisoarea noastră deschisă în favoarea siguranței IA: <https://twitter.com/elonmusk/status/554320532133650432>.
6. Erik Sofge își bate joc de știrile panice declanșate de scrisoarea noastră deschisă: „An Open Letter to Anyone Tricked into Fearing Artificial Intelligence” (*Popular Science*, 14 ianuarie 2015), online la <http://www.popsci.com/open-letter-everyone-tricked-fearing-ai>.
7. Elon Musk anunță pe tweeter donația către Institutul pentru Viitorul Vieții și către cercetătorii în siguranța IA: <https://twitter.com/elonmusk/status/555743387056226304>.
8. Mai multe despre parteneriatul pentru IA în beneficiul oamenilor și societății găsiți la <https://www.partnershiponai.org>.
9. Câteva exemple de rapoarte recente conținând opinii despre IA: <http://tinyurl.com/stanfordai>, <http://tinyurl.com/obamaAIreport>, <http://tinyurl.com/AIjobsreport>, http://standards.ieee.org/develop/indc/ead_v1.pdf, <http://tinyurl.com/roboticsmap>.
10. Între principiile care n-au fost adoptate, unul dintre preferate¹ fost următorul: „Avertisment privind conștiința: Neexistând în această privință, trebuie evitate presupunerile tranșante în punsul la întrebarea dacă o IA avansată ar avea sau ar trebui înțelesă sau sentimente.” La capătul mai multor iterații, cuvântul „conștiință” a fost înlocuit cu „experiențe subiective” – pui a primit doar 88% din voturi, cu puțin sub pragul
11. Discuție despre superintelență cu Elon Musk și alți <http://tinyurl.com/asilomarAI>.



Cuprins

Mulțumiri	7
Preludiu: Povestea Echipei Omega	11
1. Sunteți invitați la cea mai importantă dezbateră a epocii noastre...	31
2. Materia devine inteligentă	60
3. Viitorul apropiat: progrese, erori, legi, arme și joburi	95
4. Explozia inteligenței?	152
5. Consecințe: următorii 10 000 de ani	180
6. Zestrea noastră cosmică: următorul miliard de ani și mai târziu...	226
7. Obiective	278
8. Conștiința	314
Epilog: Povestea Echipei FLI	353

SERIA ȘTIINȚĂ

este coordonată de Vlad Zografi

Alte cărți de știință la Humanitas:

CARLO ROVELLI

Realitatea nu e ceea ce pare

CARLO ROVELLI

Ordinea timpului

JIM BAGGOTT

Origini:

Povestea științifică a creației

PAOLO ZELLINI

Matematica zeilor

și algoritmi oamenii

STEVEN WEINBERG

Lumea explicată:

Descoperirea științei moderne

JACOB BRONOWSKI

Ascensiunea omului

FRANS DE WAAL

Bonobo și ateul:

*În căutarea umanismului
printre primate*

Pe copertă:

© Guliver/iStockphoto/Jirsak

Noi, cu toții – nu doar oamenii de știință, industriașii și generalii –, ar trebui să ne întrebăm ce putem face pentru a profita de beneficiile viitoarei inteligențe artificiale și a evita riscurile. Aceasta e cea mai importantă dezbatere a epocii noastre, iar incitanta carte a lui Tegmark vă îndeamnă să participați la ea. —STEPHEN HAWKING

Fizicianul Max Tegmark, profesor la Institutul Tehnologic din Massachusetts, explorează posibilitățile oferite de inteligența artificială nu doar în viitorul imediat, dar și în perspectiva îndepărtată a colonizării cosmosului. *Viața 3.0* ne invită să meditam asupra celor mai delicate probleme de ordin social, militar, juridic și etic generate de dezvoltarea inteligenței artificiale, iar ideea centrală a lui Max Tegmark este că avem capacitatea și datoria morală de a influența evoluția vieții.

ISBN 978-973-50-6402-0

